DOCUMENT RESUME

ED 413 788 FL 024 919

AUTHOR	Stansfield, Charles W.; Wu, Weiping; Liu, Ching-Ching
TITLE	Listening Summary Translation Exam (LSTE) in Taiwanese (Also Known As) Minnan, Southern Fukienese, Southern Min, Xiamen, Amoy. Final Project Report.
INSTITUTION	Second Language Testing, Inc., Bethesda, MD.
PUB DATE	1997-10-09
NOTE	124p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Chinese; English; Federal Government; Interpretive Skills; *Language Tests; *Law Enforcement; *Listening Comprehension; Public Agencies; Regional Dialects; Test Reliability; Test Validity; Testing; *Translation; Uncommonly Taught Languages
IDENTIFIERS	Intelligence Gathering; *Summarization; *Taiwanese

ABSTRACT

The report details development and validation of a test designed to assess the ability to comprehend and summarize, in English, recorded conversations spoken in Taiwanese. The language and topics of the exam are representative of conversations that federal law enforcement or intelligence agencies may need to monitor in this language. The report is presented in nine sections. The first discusses the language itself and gives other relevant information on the project. The second provides a general description of the operational version of the Listening Summary Translation Exam in Taiwanese. The third describes development of the two pilot test forms, and the fourth describes development of associated data collection instruments, including the examinee background questionnaire and self-assessment instruments. The fifth section is a description of the field test sample, based on data gathered on the background questionnaire and self-assessments. Section six includes a description of the field test administration, a psychometric analysis of the field test version of the instrument, and a description of revisions made in the final version following field testing. Reliability and validity are addressed in the subsequent two sections, and the ninth section discusses the equating of the two test forms. Contains 11 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 788

Listening Summary Translation Exam (LSTE)
in
TAIWANESE
(ALSO KNOWN AS)
MINNAN
Southern Fukienese
Southern Min
Xiamen
Amoy

Final Project Report

Charles W. Stansfield
Weiping Wu
Ching-Ching Liu

Second Language Testing, Inc.
10704 Mist Haven Terrace
N. Bethesda, MD 20852

September 30, 1996
Revision submitted October 9, 1997

FL024919

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Weiping Wu
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TABLE OF CONTENTS

Acknowledgments		1
Abstract		2
1.0	Background Information	3
1.1	The Language	3
1.1.1	The Name	3
1.1.2	Language communities	3
1.1.3	History	4
1.1.4	Role in society	4
1.1.5	Dialects	5
1.2	Origins of This Project	5
1.3	Project Staff	6
2.0	Description of the LSTE-Taiwanese	8
2.1	Multiple-choice Section	8
2.1.1	Format	8
2.1.2	Test Taking	9
2.1.3	Scoring Procedures	10
2.2	Summary Translation Section	10
2.2.1	Format	10
2.2.2	Test Taking	10
2.2.3	Scoring Procedures	11
3.0	Development of the LSTE-Taiwanese	13
3.1	Development of Conversations	13
3.2	Exam Forms	16
3.3	Exam Tapes	17
3.4	Other Test Materials	17
3.5	Development of Materials Used to Score the Summary Translations	17
3.6	Field Test Administration.	18
3.7	Development of the Accuracy Guidelines	19
3.8	Development of the Expression Rating Scale	20
3.9	Development of the Expression Scale Self-Instructional Rater Training Materials	20
4.0	Development of Ancillary Data Collection Instrumentation	22
4.1	Development of Self Assessment Questionnaires	22
4.2	The Self-Assessment of English Writing Ability (SA-EW)	22
4.3	The Listening Comprehension Global Self-Assessment Questionnaire (SA-LC)	24
4.4	The Self Assessment of Summary Translation Ability	25
5.0	Description of the Field Test Sample	27
5.1	Background Questionnaire	27
5.2	Characteristics of Field Test Sample	28
5.2.1	Data collection locations	28
5.2.2	Ethnic background and nationality	28
5.2.3	Mother's Place of Birth	28

5.2.4 Father's Place of Birth 28
5.2.5 Individuals from whom Minnan was learned 29
5.2.6 Current Minnan interlocutors 29
5.2.7 Amount of time Minnan is spoken in a typical week 29
5.2.8 Age 30
5.2.9 Age when Minnan was learned 30
5.2.10 Time lived in Minnan-speaking area 30
5.3 Results of Self-Assessment Questionnaires 30
5.3.1 Self Assessment of English Writing Ability 30
5.3.2 Self Assessment of Listening Comprehension in Minnan 31
5.3.3 Self-Assessment of Summary Translation Ability 31
5.3.4 Total Summary Translation Ability 33
5.4 Summary 34

6.0 Field Testing and Revision 36
6.1 Field Test Administration 36
6.2 Results of Field Testing: Multiple-choice Portions 36
6.3 Results of Field Testing: Summary Writing 37

7.0 Reliability 39
7.1 Reliability of the Accuracy Score 39
7.1.1 Reliability of the Multiple-choice section 39
7.1.2 Accuracy Score: Summaries 40
7.1.3 Reliability of the Total Accuracy Score 42
7.2 Reliability of the Expression Score 42

8.0 Validity 47
8.1 Content Validity 47
8.2 Criterion-related Validity 48
8.2.1 Interrelationships between Test Scores 49
8.2.2 Relationships between the LSTE-Minnan Accuracy Scores and the Self-Assessments 51
8.2.3 Summary of Evidence for the validity of the Accuracy scores 55
8.2.4 Relationships between the LSTE-Minnan Expression Score and the Self-Assessments 55

9.0 Equating LSTE-Minnan Accuracy Scores 57
9.1 Equating the Forms 57
9.2 Construction of the Summary Accuracy Scale 58
9.2.1 Overview 58
9.2.2 The Selection of the Criterion Variable 59
9.2.3 Outliers Detected and Removed 61
9.2.4 Effects of Removing Outliers 62
9.2.5 Development of Raw Score to Summary Accuracy Score Conversion Tables 62
9.3 Further Analyses of the SAS Scores 63
9.4 The Final Accuracy Rating 64
9.5 Using the Multiple-Choice Section as a "Screen" 65

References 67

List of Appendices

A....Multiple-Choice Section Test Booklet (Selected Pages)
B....Summary Section Test Booklet (Selected Pages)
C....Expression Scoring Guide
D....Self-Assessment of English Writing Ability
E....Self-Assessment of Listening Comprehension
F....Self-Assessment of Summary Translation Ability
G....Test-Taker Background Questionnaire
H....LSTE-Minnan Exam Feedback Questionnaire (Pre-Field Testing)
I....Instructions to Field Test Administrators
J....Interpretation of 0-5 Final Accuracy Rating Scale
K....Score Conversion Table: Form B Raw Score to Form A Raw Score
L....Summary Accuracy Scale (ILR-Based) Conversion Tables
M....Scatterplots Depicting Predicted and Observed SAS Scores from Test Scores
N....Abbreviations and Abbreviation Equivalencies

Acknowledgments

While there are many individuals and organizations that contributed to this project, we wish to express our special appreciation to those that follow. The Center for the Advancement of Language Learning provided funding the project. Mrs. Marijke Cascallar of the Federal Bureau of Investigation, the lead agency for this project, wrote the statement of work reflecting the FBI's operational requirements. She also provided general oversight and was helpful in solving problems and arranging for the cooperation of FBI field offices. The Center for Applied Linguistics (CAL) graciously allowed us to utilize its facilities on several occasions for staff meetings, consultant meetings, and communications. Dr. Dorry Kenyon of CAL performed the statistical analyses and equating, and reviewed the description of the equating in this final report. Ms. Barbara Hicks of the Center for Equity and Excellence in Education at George Washington University evaluated all the examinee responses produced in Section B of the test, and developed the rater training materials for the Expression score.

A number of other individuals played a lesser, although still important role, in the project. These include the following. Dr. Mary Lee Scott of Brigham Young University, Dr. Qinghai Chen of the University of Michigan, Dr. Ying-Che Li of the University of Hawaii, Dr. Cornelius Kubler of Williams College, and Kairen Zhuang of Columbia University, performed a variety of advisory and quality control tasks as external consultants to the project. Dr. Ronald Walton, Deputy Director of the National Foreign Language Center, showed great interest in the project, offered informal guidance, and suggested appropriate contacts. Dr. Pardee Lowe of the CIA Language School, Dr. Mats Oskarsson of the University of Gotenberg, and Dr. Kenneth Wilson of Educational Testing Service reviewed the self-assessment questionnaires. Bryant Rogers and other CALL provided useful advice upon request and carried out administrative oversight tasks in connection with this contract. Dr. John L.D. Clark, former Dean of Testing, Research and Evaluation at the Defense Language Institute, reviewed and commented on the final report, as did Dr. Kenneth Wilson.

Finally, we would like to thank the Center for Advancement of Language Learning (CALL) for funding our proposal to carry out this project. We appreciate their confidence in us and we hope that the work depicted herein will serve as an example to other Government agencies of the kind of work that can be done by competent contractors.

Abstract

This final report is divided into nine narrative sections. The first section discusses the language which is the subject of the test and gives other relevant background information on the project. The second section provides a general description of the operational version of the Listening Summary Translation Exam (LSTE) in Taiwanese.[1] The third describes the development of the two pilot forms of the test. The fourth section describes the development of associated data collection instruments, including the examinee background questionnaire and the self assessment instruments. The fifth section is a description of the field test sample based on the data gathered on the background questionnaire and the self-assessment instruments. Section six includes a description of the field test administration, a psychometric analysis of the field test version of the instrument, and a description of the revisions made in the final verions following field testing. The seventh section discusses the reliability of the LSTE-Minnan, including subtests and subscores, based on a reanalysis of data following the deletion of items in the final version of the tests. Section eight uses a correlational approach to present and discuss the validity of the test. The ninth and final section discusses the equating of the two forms. The report also includes a References section. Appropriate appendices supplement the narrative.

---

[1]The original RFP for this project called for an LSTE in Fukienese. Linguistically speaking, this is an imprecise term. After the project began, it was determined that the language of the test would be Minnan, which is spoken in the southern part of Fujian province. The dialect ultimately tested is Amoy, which is informally called Taiwanese. As a result, the terms used to refer to the language of the test vary in this report and on test materials, according to the context. However, generally this technical report uses the term Minnan to refer to the language of the test. Literally, Minnan means Southern Min in Mandarin Chinese.

1.0 Background Information

1.1 The Language

1.1.1 The Name

The original RFP called for a test of "Fukienese." Subsequently, it was determined that the language of the test would be Minnan, which is the principal language of Fukien Province. Later is was determined that the dialect to be used on the test should be Amoy, which is the principal dialect of Taiwan, and which normally is referred to in English as "Taiwanese." More detailed information about the names that are used to refer to the language and the dialect follow.

Many terms are used to refer to the language of this test. A very common term is Minnan. Min3nan2 is the Pinyin transliteration of the Mandarin Chinese word that means "southern Fujian (Province)". An alternate term, Hokkien, represents the Minnan pronunciation of "Fujian." The terms "Fukien" and "Fukienese" are the Anglicized versions of "Fujian" and "Fujian language" (or speech) respectively. The language is sometimes referred to as Xiamese or Taiwanese because the Xiamen dialect is considered its most representative. It is spoken, though with variations, in most parts of Taiwan. It is called the Amoy (or Amoi) language. Amoy (an established English word) is the way Min3nan2 (literally "Southern Fujian") speakers pronounce the word "Xiamen." When used without the modifiers "nan2" (southern) or "bei3" (northern), Fukienese, Fukien, Hokkien, or even Fujian just mean "Minnan," i.e., the southern Fujian language. By comparison, the Minbei (Northern Min) language is less important than Minnan in number of speakers, either within the boundaries of Fujian Province or beyond it, although recently some number of Minbei speakers have left Mainland China for the US. Because Minnan is the preferred term for this language, in this report we refer to it as such. However, in order to facilitate its identification to US Government employees who may not have a background in linguistics we use the term Taiwanese on the test booklets and other materials used by examinees and raters.

1.1.2 Language communities

Out of 49 million total speakers (1991), 25.73 million are in China, excluding Taiwan and Hong Kong (1984). When all three areas are considered, the speakers combined account for 3% of the Han Chinese population. Within China Minnan is spoken in the southern part of Fujian Province (Prov.), the eastern part of Guangdong Prov., parts of Hainan Prov., the southern areas of Zhejiang Prov. and Jiangxi Prov. It is spoken in most parts of Taiwan, with speakers numbering 14.18 million. There are .54 million Minnan speakers in Hong Kong, 1.95 million in Malaysia, 1.17 million in Singapore, 1.08 million in Thailand, .70 million in Indonesia, .49

million in the Philippines, .01 million in Brunei. Minnan is spoken also in other countries, although exact statistics are not yet available.

1.1.3 History

Languages in Fujian (Minnan included, or course) have a relatively long history. By the turn of the Sui (598 - 617 AD) and Tang (618 - 906 AD) Dynasties most of them had taken shape already. Evidence includes the existence in today's Fujian languages of certain pronunciations and lexical items of Middle and even Early Chinese. Minnan in particular is the linear descendent of the standard or prestige language in the Middle Chinese period, whose phonetic features are recorded in *Qielyun4*, a rhyme book believed to have been compiled during the late Sui years and widely adopted by the Tang poets. In fact, the Tang poems extant today would sound perfectly fitted to the required rhyme patterns if intoned in the Minnan language. Fujian Province's geographical features and corresponding relative isolation from other parts of China account for the largely stable evolution of its languages.

1.1.4 Role in society

Although used extensively by native speakers on many occasions, Minnan is not (or is not supposed to be) the language of instruction in schools, at least in larger towns and cities. Nor is it the language used on official occasions. However, in rural areas Minnan is used more frequently.

Local radio stations in the southern parts of Fujian as well as Fujian Provincial Radio have programs in Minnan, and the latter has Minnan programs specially designed for listeners in Taiwan. The Beijing-based Central People's Radio also has special programs in Minnan meant for listeners in Taiwan, Hong Kong, Macau, and foreign countries. Taiwan has local radio stations broadcasting in Minnan, and special Minnan programs meant for listeners in mainland China, as well. The British Broadcasting Corporation (BBC) has programs in Minnan, as does the NHK in Tokyo.

More and more Minnan speakers are becoming fluent in Putonghua (Mandarin) because of the educational efforts made by governments both in mainland China and Taiwan. This tendency has been so strong that in larger cities younger people, especially the more educated ones, are becoming less proficient in their native Minnan than their parents. Yet for certain practical purposes such as language maintenance, Minnan is still being taught to non-speakers. One instance can be found in Singapore, where the *Straits Times* carries advertisements by various language centers to recruit students into Minnan courses.

It should be noted that Minnan plays a more important role in Taiwan than it does in mainland China or Singapore. In Taiwan,

Minnan is referred to as "tai2yu3," the Taiwanese language. Not only is Minnan greatly influential in popular culture and entertainment through the mass media of television and radio, it has a place in political life, also. For example, some Representatives to the Congress, most of them belonging to opposition parties, prefer to speak and debate in Minnan. Furthermore, there has been some discussion in the Congress as to whether Minnan should be conferred equal status with Mandarin.

1.1.5 Dialects

To be linguistically exact, Minnan is not one language but a group of dialects. Variant pronunciations alone are enough to cause difficulty in mutual understanding, especially when they represent different dialects. Yet within the same dialect, speakers of different local variants understand each other with relative ease.

The major dialects are Xiamen, Quanzhou, and Zhangzhou. 70-80% of the Taiwanese people can trace their family origins to these areas and dialects. However, the Xiamen dialect (including its sub-variants) is most widely used in such media as radio broadcasts. Although the Xiamen variant cannot be said to possess the prestige status that the Guangzhou variant does within Cantonese dialects, it is accepted most readily when speakers of different variants talk together. The Xiamen subdialect used in Taiwan, called Amoy, is the dialect used on this test. It is noteworthy that initially SLTI staff planned to include different Minnan dialects on the test. However, feedback from the FBI indicated that Amoy should be the sole dialect included.

1.2 Origins of This Project

The LSTE format was originally developed by Charles W. Stansfield and his associates at the Center for Applied Linguistics. The first Listening Summary Translation Exam (LSTE) was in Spanish. Both the Spanish and Minnan tests authentically reflect FBI operational tasks and requirements and are designed to meet the Bureau's operational needs. The LSTE-Spanish was a response to a statement of work prepared by the FBI. Following its development, it was implemented successfully by the FBI. The FBI then requested that CALL fund the development of similar tests in other languages. During May 1994, the Center for Advancement of Language Learning (CALL) distributed a request for proposals (RFP) for Listening Summary Translation Exams (LSTE) in 10 languages and dialects. SLTI responded to the RFP. Due to limitations in funds, only three languages were funded: Arabic, Mandarin, and Southern Fukienese (hereafter normally referred to as Minnan or Taiwanese). The Arabic and Mandarin tests were awarded to the University of Illinois. The test of Fukienese (ultimately determined to be

Taiwanese/Minnan) was awarded to SLTI. The Minnan project was funded at the end of September 1994. Work on the project began in November 1994. The project was completed on September 30, 1996. Some time later, revisions were requested on the final report. These revisions were completed in March 1997, with the submission of this final report.

1.3 Project Staff

The project was directed by Charles W. Stansfield, who also served as project director. Dr. Stansfield has over 25 years experience in the language testing field. He has developed important national tests in over a dozen foreign languages. He has made many contributions to the language testing field, and served as the founding President of the International Language Testing Association (ILTA). Prior to founding SLTI, Dr. Stansfield was director of the Division of Foreign Language Education and Testing at the Center for Applied Linguistics. In this position, he also served as director of the ERIC Clearinghouse on Languages and Linguistics. In response to the statement of work, Dr. Stansfield conceptualized and was involved in all phases of the project. He served as one of two raters for the Expression score and is the principal author of this final report.

Dr. Weiping Wu served as project coordinator. Dr. Wu is a native speaker of Minnan. He holds a Ph. D. in Linguistics from Georgetown University. His specializations are forensic linguistics and language testing. For the past five years he has worked as a test developer at the Center for Applied Linguistics (CAL). In that capacity he has participated in the development of tests of Mandarin and Cantonese (which he also speaks), Hindi, and Russian. Throughout the project, Dr. Wu worked approximately half time for SLTI and half time for CAL. Dr. Wu was involved in all phases of the project. Among other tasks, he wrote test items, translated the conversations to English, scheduled the work of others, and provided second ratings using the Accuracy checklists.

Ms. Ching-Ching Liu served as project assistant. Originally, from Taiwan, she is a native speaker of Minnan, Hakka, and Mandarin. She has a Master's degree in Linguistics from the University of Delaware. She has also worked at the Center for Applied Linguistics, where she helped produce the Mandarin Simulated Oral Proficiency Interview (SOPI) Rater Training Kit. During the course of the project, she worked approximately half time for SLTI while she taught Chinese half time at the American University. Ms. Liu was involved in all phases of the project. Among other tasks, she wrote test items, transcribed the conversations, rated the summaries for Accuracy, and drafted the Guidelines for Scoring for Accuracy.

Dr. Dorry M. Kenyon served as a statistical consultant to the

project, and was responsible for all statistical analyses. He holds Ph.D. in Educational Measurement and Evaluation from the University of Maryland. He is director of the Division of Foreign Language Education and Testing at the Center for Applied Linguistics.

Ms. Barbara Hicks served as a consultant at the end of the project. She drafted the rater training materials for the Expression score. She also served as one of the two Expression raters. She is currently completing a Master's degree in Applied Linguistics at Georgetown University and is on the staff of the Center for Equity and Excellence in Education at George Washington University, where she carries out projects relating to educational testing.

A number of other individuals served as consultants or assisted the project in various ways. They are mentioned in the body of the report or in footnotes that associate them with a particular activity in the test development process.

2.0 Description of the LSTE-Taiwanese

The Listening Summary Translation Exam in Taiwanese (LSTE-Taiwanese) is designed to assess the ability to comprehend and summarize in written English recorded conversations spoken in Taiwanese. The language and topics of the exam are representative of conversations that the FBI and other federal law enforcement or intelligence agencies may have need to monitor in this language.

The LSTE-Taiwanese consists of two subtests. The first contains 50 multiple-choice items based on twelve recorded conversations. This subtest is referred to in this part of the report as the multiple-choice section or Section A. The second subtest requires examinees to write summaries of three recorded conversations. This subtest is referred to in this report as the Summary Translation section or Section B.[2] A separate test booklet for each section contains instructions, example items, and test items. A master tape for each section contains the general introduction to the exam,[3] instructions, example items, and recorded conversations. The LSTE-Taiwanese exists in two forms that are generally parallel in content, item difficulty, format, and length.

2.1 Multiple-choice Section

This section of the report describes the format, test taking, and scoring procedures for the multiple-choice section of the LSTE-Taiwanese. While the multiple-choice section contributes to the total score, it is also used as a screeing test. The screening function is discussed later in section 9.5 of the report.

2.1.1 Format

There are 50 items in the multiple-choice section, based on twelve recorded conversations. These conversations simulate exchanges regarding drug deals, fraud, terrorism, gambling, illegal immigration, and military and political affairs. Because they are unscripted, the conversations manifest all of the characteristics

---

[2]Section A was incorrectly called Part A during the recording of the test. Because of the subsequent inavailability of the speaker who recorded the general instructions and introduction to the test, we had to continue to refer to the two portions of the test as Part A and Part B. However, in this technical report on the project, we will refer to the two portions as Section A and Section B.

[3]Examinees are informed that they will hear brief conversations involving two people, and that the age, sex, and regional accent of the speakers will vary.

of natural speech, including hesitations, false starts, repetitions, interruptions, overlapping of speakers, misunderstandings, requests for clarification, etc.

The test items vary in purpose: some of them assess comprehension of specific details such as dates, times, locations, etc., while others require the examinee to infer the relationship of the speakers, their emotional reactions to the messages conveyed, and possible actions to follow from the conversations.

A test booklet contains instructions, example items, explanations, and the test items themselves. Appendix A contains selected portions of a test booklet for the multiple-choice section, including the cover page, instructions, and example items.

2.1.2 Test Taking

Each examinee receives a multiple-choice section test booklet, a machine scoreable answer sheet, and two no. 2 pencils. Examinees listen to the test instructions on the tape, and read along in their test booklets when instructed to do so.

Examinees are informed that they will hear a series of conversations, some of which are related to each other. In this section, each conversation is presented only once.[4] Examinees are given a block of time before hearing a given conversation to scan the questions and options pertaining to that particular conversation.[5] By scanning the items before hearing the conversation, they have an idea of what type of information to listen for.[6]

---

[4]In an actual work setting, the listener would be able to replay the conversation as many times as needed. However, repeated playing of taped conversations indicates a lack ability to understand the conversation. In addition, it reduces the productivity of the listener in that he or she can screen fewer tapes in a given amount of time if conversations must be repeatedly replayed in order to be understood. Therefore, for purposes of a test, allowing the examinee to hear the conversation only once is believed to be a practice that contributes to the predictive validity of the instrument.

[5]There are four to six items for every conversation. Examinees are given from four seconds to scan each item. Thus, if the conversation is followed by six items, examinees are given 24 seconds to scan the items before listening to the conversation.

[6]All of the points tested are considered important. In an actual on-the-job work setting, the person listening to the tape would know what information is important and what isn't.

As they listen to the conversation, examinees may read the items again. They are cautioned not to be distracted by slang or phrases that are unfamiliar to them. Instead, they are to concentrate on extracting only the information needed to answer the questions.

After listening to the conversation, examinees are given 15 seconds per item to read the stem and options, to decide on the answer, and to transfer the answer to the answer sheet. The entire process is paced by the tape so that every 15 seconds the examinee is told to go on to the next question.

The multiple-choice section lasts approximately 55 minutes. The exact figures are 54 minutes and 36 seconds for Form A, and 56 minutes and 38 seconds for Form B.

2.1.3 Scoring Procedures

Examinees record their responses to the multiple-choice section of the LSTE-Taiwanese on answer sheets that are scored by machine. The score on this section is the number of answers correct. The maximum possible score is 50.

2.2 Summary Translation Section

The section that follows describes the format of the Summary Translation portion of the LSTE-Taiwanese, as well as the test taking and scoring procedures for the test. The Summary Translation portion is also called Section B of the test.

2.2.1 Format

In the Summary Translation portion, examinees are required to summarize three conversations, which increase in length (from approximately one to three minutes) and in sophistication of vocabulary. The conversations are similar to those in the multiple-choice portion. However, in Section B examinees hear each conversation twice, and they are permitted to take notes on the content of the conversation.

The Summary Translation test booklet contains instructions, space for taking notes and writing a summary of an example conversation, observations regarding the example summary, and space for taking notes and writing summaries of the remaining conversations. (Appendix B contains selected portions of the test booklet for Summary Translation, including the cover page, instructions, an example summary, and an analysis of the example summary.)

2.2.2 Test Taking

In Section B, examinees hear each conversation twice. They

take notes as they listen to the conversation, and then write a summary in English using the information in their notes. Both the notes and summaries are written in the test booklet.

Examinees are allotted from three to 13 minutes to write summaries of the conversations, depending on the length of the conversation. Before beginning a particular summary, examinees are informed of how much time they will be given. They are also advised when there is one minute remaining to complete the summary.

The instructions for Section B are designed to train the examinee to write an effective summary, if the examinee is competent to do so. Examinees are told what kind of information should be present in an effective summary, including the overall topic of the conversation, and supporting details including names, dates, times, places, or amounts. As conversations vary in the amount of concrete information they contain, examinees are cautioned to make sure they identify the general topic and primary supporting points of more abstract conversations. They are instructed to include as much detail as possible in the summary. However, they are to include only information they have gleaned from the conversation, and not to add any of their own assumptions or inferences. The use of a bilingual dictionary (and similar reference materials) is permitted on Section B of the test.

The duration of Section B is approximately 52 minutes. The exact figures are 51 minutes and 26 seconds for Form A, and 53 minutes and 24 seconds for Form B.

2.2.3 Scoring Procedures

Examinees receive two scores for Section B: one for Accuracy and the other for written Expression. Both are assessed by a trained rater.

Accuracy is scored by the rater through the use of a checklist that identifies the callers, the main topic, and key and supporting points in the conversation. As the rater reads a summary, he or she checks off those items on the list which the examinee has reported accurately; one point is awarded for each key and supporting point. Although the wording of the summary does not have to match exactly that of the checklist, it is important that the information be provided in the appropriate context. Because the content of the conversation is broken down into items of information on the checklist, an examinee can receive credit for each item that is accurately reported, even if other items are omitted or misunderstood. The Accuracy score is the sum of the points awarded for each of the three conversations. The maximum number of points for Accuracy on Form A of the LSTE-Taiwanese is 50; on Form B it is 56.

Expression is scored by the rater through an evaluation of the written summary for correct grammar, spelling, punctuation, and syntax, precision of vocabulary, and organization. The principal criterion is communicative effectiveness of the English employed. An inability to communicate the intended information generates the lowest rating on the Expression scale. This written summary is evaluated according to the Expression Scoring Guide (see Appendix C). For each of the three summaries, the examinee is awarded either a "Deficient" (= 1 point), "Functional" (= 2 points), or "Competent" (= 3 points), or "Native" (= 4 points). The total Expression score is the average of the Expression scores on the three summaries.[7] Once the average is computed, a final rating is awarded as follows:

Average Expression Score	Final Rating
1, 1.33	Deficient
1.5, 1.67, 2.00, 2.33	Functional
2.5, 2.67, 3.00, 3.33	Competent
3.5 3.67, 4.00	Native

The Accuracy and Expression scores on the LSTE-Taiwanese Summary section are always kept separate. However, a total score for Accuracy (TOTACC) on the LSTE-Taiwanese is awarded by adding the raw score on the multiple-choice Section and the Accuracy score on the Summary Translation Section together. The maximum Total Accuracy score obtainable on Form A of the LSTE-Taiwanese is 100; on Form B it is 106.

---

[7]If one of the summaries is so short (e.g., a few words or a single sentence) that it can not be rated for English Expression, it is designated "Unratable," and is not counted in the final Expression score. In this case, the other two summaries are averaged and the average becomes the final Expression score. At least two summaries must be ratable in order for an Expression score to be obtained.

3.0 Development of the LSTE-Taiwanese

This section describes how the two pilot forms of the LSTE-Taiwanese were developed. The method of developing the simulated conversations, the preparation of examination materials, and the pilot study scoring methods are described.

3.1 Development of Conversations

Because the LSTE-Taiwanese is designed to be used in occupational settings, project staff felt that it was important that the conversations used on the test be as authentic as possible. For this reason, staff obtained information that influenced the nature of authentic conversations from a variety of sources. These include taped conversations provided by the FBI, interviews with private contractors who listen to and transcribe tapes provided by the FBI and other law enforcement agencies on a daily basis, and interviews (telephonic and face-to-face) with FBI staff that listen to such conversations. This approach to creating authentic conversations was used in the development of the LSTE-Spanish and is analyzed and validated in Scott, Stansfield, & Kenyon (1996).

Originally we planned to use taped conversations from adjudicated cases provided by the FBI to inform the development of the conversations on the LSTE-Taiwanese. Several months after the project began the FBI informed SLTI that this would not be possible. However, we did obtain tapes in Taiwanese, Mandarin and Cantonese from a federal court in Boston involving a Drug Enforcement Administration case. Subsequently, we were able to obtain additional tapes from the FBI.

Summary of linguistic features. In preparation for the creation of conversations, we conducted an informal analysis of the DEA tapes in order to identify the general characteristics of the conversations that might be monitored by law enforcement agencies. The analysis included identification of frequent topics, tone, and use of nicknames, colloquial expressions, and code words. We then prepared a summary of the general characteristics we discovered. We also interviewed a manager of an FBI contractor in New York City who is engaged in this type of translation activity. In addition, we developed a number of brief scenarios outlining the gist of conversations to be used for the LSTE-Taiwanese.

Telephone questionnaire. In order to systematically gain information from staff listening to tapes at FBI field offices, SLTI staff prepared a questionnaire that guided telephonic interviews. A draft questionnaire was sent to the FBI for review in December 1994 and following revisions, the final version was completed in January 1995. The questionnaire dealt with the language background of the linguist, the age, sex, and background of participants in audited conversations, the nature of the

conversations in terms of topics, type, and tone of language used in the conversations, the sources, types, and frequencies of conversations that they listen to, the general content areas, and the topics within each area. This questionnaire was used when interviewing FBI staff and contract linguists at different field offices. Sometimes the interviews were conducted in English, sometimes in Mandarin, sometimes in Minnan, and sometimes in a combination of the three languages. The SLTI staff member who conducted the interview made notes on each.

Because of security concerns as well as internal FBI policy and practice, those interviewed were not always able to fully respond to our questions. However, excellent cooperation was received from the language supervisor at the FBI Los Angeles field office, where contract linguists working with Minnan provided extensive information and examples, and subsequently reviewed and critiqued a draft version of the entire test. For security reasons, the names of interviewees were generally not provided to SLTI. Also, in no case were the interviewees allowed to discuss anything related to foreign counter-intelligence (FCI) work.

Revised summary of linguistic features. SLTI staff and consultants met with FBI staff to discuss the general characteristics of monitored conversations, the scenarios which had been developed to that point, and the exam format and scoring. As a result of this meeting, the original summary of linguistic features was revised and expanded with information obtained from FBI staff.

Consultants. Because of the need to gather more explicit information on FCI topics and language, and the inability of FBI staff to discuss these matters with the test development team, SLTI contracted as consultants two Sinologists who are political science professors with considerable knowledge of sensitive issues.[8] Based on the information they provided, we were able construct scenarios in the FCI area, which were judged to be realistic by FBI staff.

Taxonomy and scenarios. Based on all of the information gathered, a taxonomy containing 37 topics and tasks (speech functions) was developed. This taxonomy was also reviewed by the FBI and refined based on comments. Subsequently, draft scenarios

---

[8] The consultants were Professor Lin Chongping of Georgetown University and the American Enterprise Institute, and Ralph Clough of the School of Advanced International Studies at Johns Hopkins University (a former foreign service officer assigned to Taiwan). Both were well informed about military and political matters involving the US, Taiwan, and the PRC. Indeed, the weeks following our conversations showed the test developers that the consultants had successfully predicted the diplomatic turn of events involving these countries.

of conversations were developed to match each topic and task. In this way, it became possible to inspect the content objectives (the topics and tasks in the taxonomy) and the way it was proposed that each objective would be tested. The taxonomy and draft tasks was submitted to FBI Headquarters. There it was critiqued by staff in the Language Services Unit, and it was forwarded to field offices with Minnan-speaking staff. Staff were asked to rate each objective and proposed conversation on a five point scale in terms of its frequency of occurrence and difficulty. The written evaluations of individual reviewers were returned to and tallied by SLTI.

The analysis indicated that most proposed conversations were viewed as frequently occurring, thereby indicating their validity for inclusion on this occupational test. The conversations rated as frequently occurring were also rated as easy to work with. However, a few were viewed as rarely occurring and not easy to work with. These rarely-occurring, more difficult conversations dealt with matters related to foreign counter-intelligence (FCI) work. Still, SLTI and the FBI felt it important to include a number of FCI conversations on the test. Such conversations increase the range of proficiency assessed by the instrument, and they make the test useful in the selection of a wider number of occupational specialties within the FBI and the US Government at large.

Selection and Training of Actors. Following further revisions and the writing of some additional FCI scenarios, SLTI staff interviewed 13 native speakers of Minnan who were willing to serve as actors in the recording of the conversations. Of these, we determined that nine individuals (seven of those interviewed plus two SLTI staff members) had the language proficiency and personal skills necessary to improvise the conversations based on the scenarios. The actors varied in age and spoke two dialects of Minnan: Amoy and Chaozhou. Six were male and three were female. Seven of the speakers spoke Amoy. SLTI staff trained the actors used in each taping session. Training involved a review of the general characteristics of monitored conversations followed by practice tapings. The actors were encouraged to speak naturally and to use slang, regionalisms, or even vulgarities that would be appropriate in a given situation.

Recording conversations. After reviewing the scenario for a given conversation, the actors agreed on code words and basic content, and rehearsed the conversation briefly several times face-to-face. One called the other on a phone (both phones were different extensions located at different desks at an office but were located in the same work area) and carried out the conversation by phone. The conversations were taped using a recording device attached to one of the phones, thus simulating as closely as possible conditions under which conversations are often monitored by the Bureau. A conversation was re-taped as many times as needed until it was determined to be wholly authentic by SLTI

and FBI staff. An FBI linguist of the Washington, DC Field Office, was present at the initial recording sessions in order to provide feedback to the actors on the authenticity and acceptability of the conversations as they were being taped. Due to lack of vocabulary and complex sentence structures, not all actors proved capable of carrying out the FCI conversations. As a result, some taping sessions had to be rescheduled so that another speaker could be brought in and trained to carry out an FCI type conversation.

A total of 36 different conversations were taped over a number of recording sessions. Each test tape contains all of the speakers, with the result that a variety of voices are represented on each test form.

Review of preliminary conversations. A tape was constructed based on the conversations recorded and sent to the FBI field office in Los Angeles. There, two Minnan speaking contract linguists listened to each conversation and evaluated it using a questionnaire prepared for that purpose by SLTI. The questionnaire dealt with the authenticity of the language used in the conversations as well as the clarity, rate of speech, etc. Most conversations received high marks in this review. However, those conversations that involved the Chaozhou dialect of Minnan were considered as generally unintelligible by these linguists. The linguists also questioned the appropriateness of including this dialect on the test. As a result, a decision was made by FBI Headquarters to remove all conversations with Chaozhou dialect from the test. This reduced the total number of speakers used on the final forms to seven, four males and three females.

3.2 Exam Forms

SLTI staff and consultants wrote multiple-choice items based on a number of the recorded conversations. The items were designed to assess the understanding of specific information and the ability to make inferences based on the information presented in the conversations.[9]

Parallel forms of the LSTE-Taiwanese were constructed so as to ensure a similar distribution of the number of conversations (for each form, 12 in the multiple-choice section and 3 in the Summary section), length of conversations, the sex of the speakers, and the number of multiple-choice items which had been developed (57 items for the pretest versions, which became 50 items in the final

---

[9]The items in the Multiple Choice part differed in this aspect from the instructions given in the Summary Writing part, which cautioned the examinee not to insert his or her own inferences in writing the summary, but to report only the information presented in the conversation.

versions). After developing the answer key for the multiple-choice portion of each form, we made changes in the ordering of the options to ensure equal distribution of correct answers across the four choices A, B, C, and D. More conversations and items than would be needed on the final versions were prepared, so that only those that functioned most effectively could be retained.

3.3 Exam Tapes

After organizing the conversations and items into parallel forms, we prepared scripts for the narration of each form. The scripts included a general description of the exam, instructions for filling out the machine scoreable answer sheet and test booklet, example items and explanations, multiple-choice and summary item numbers, and instructions to the recording engineer for placement of the recorded conversations.

SLTI worked with a professional recording studio, Lion and Fox, Inc., to edit and assemble the conversations from individually recorded cassette tapes into the two test forms. The narration of the forms was recorded in the studio by a professional radio announcer who works for a local public radio station. Subsequently, the narration and conversations for each form were merged on to a master tape. At this time the pauses before each conversation and between items were inserted. Cassette copies for use in the pretesting were made from the master tape.

3.4 Other Test Materials

SLTI also prepared test booklets for each form of the LSTE-Taiwanese (as described in section 1 of Section of the report). In addition, we prepared detailed directions for test administrators on how the administer the tests, the background questionnaire, and the various self-assessments. The test administration instructions included information regarding: 1) test security, 2) assembling test materials, 3) selecting and arranging for a suitable testing site, 4) equipment, 5) administering the test (including the timing of sections), and 6) procedures to follow in returning test materials.

3.5 Development of Materials Used to Score the Summary Translations

Scoring procedures for the LSTE-Taiwanese are modeled on the LSTE-Spanish. The scoring of the multiple-choice section of the test were objective and straight-forward; since there was only one correct answer to each question. For the Summary Translation section, however, we wanted the scoring procedures to focus on the examinees ability to record important information that the conversation contained. Consequently, we devised a plan to

identify the important points in the Summary Translation section conversations.

In order to do this, we wrote a summary of each of the conversations by listening to the conversation several times, stopping and re-playing the tape as often as needed in order to capture as much detail as possible. We also transcribed the conversation using traditional Mandarin Chinese characters and we then translated the transcription into English. Referring to the tape, the transcription, and the translation, we constructed a checklist of important points mentioned in a good summary for each conversation. FBI language specialists and three external consultants then read these sample good summaries and the checklists to verify that the checklists included all important and appropriate information. Once the checklists were validated by the FBI, they were considered ready for use in the field test administration.

3.6 Field Test Administration.

The tests were administered at three sites: the University of Maryland, the University of California at Berkeley and the University of California at Davis.[10] These sites were selected because we knew that substantial numbers of Minnan speakers were located there, and because we were able to enlist the cooperation of our colleagues at these universities for cooperation and assistance in recruiting field test examinees. All field test data were gathered between late February and early May, 1996.

At each site, we contracted with a test administrator to recruit the examinees, to obtain space, and to administer the test on two different occasions approximately one week apart. Examinees were paid an honorarium for taking the test, and were paid again for taking the second form of the test, if they so desired. Over half of the examinees returned to take the second form of the test.

The order of administration of the forms was counter-balanced, so that approximately half of the examinees took Form A first and half took Form B first.

Following the administration, all test booklets, scannable answer sheets, administration instructions, and questionnaires were

---

[10]At the University of Maryland, Dr. Scott McGinnis, current national President of the Chinese Language Teachers Association, arranged for the administration and recruited examinees. Dr. Tim Xie of the University of California at Davis and Mr. Theron Stanford of the University of California at Berkeley performed similar duties at the their institutions. We are grateful for their contributions to this project.

returned to SLTI via Federal Express. Subsequently, honoraria were sent to each examinee and test administrator.

Once the test materials were received by SLTI, the NCS answer sheets were scanned into a database using an NCS Sentry 3,000 scanner. (In the operational program, any answer sheet can be used with the corresponding scanner.) In addition, examinee responses to Section B, the summary writing portion of the test, were scored by two trained raters. This scoring is discussed in more detail in section 3.8.

3.7 Development of the Accuracy Guidelines

Section 3.5 above describes the development of the Accuracy Checklists prior to field testing. Following the field testing, we revised the checklists and we developed a training document to make the scoring as objective as possible. This document, the *Guidelines for Scoring Each Point on the Accuracy Checklist*, provides guidance and scoring criteria for each point on the checklist. The Guidelines were carefully developed to answer the vast majority of questions a rater might have when using a checklist.

In order to develop the Guidelines, SLTI test development staff followed a specific procedure. Two Minnan-speaking test developers made an outline of the Guidelines following the scoring of a limited number of summaries with the Checklists. The outline established the basic format: sample good summary, the scoring point (the piece of information to be included), an explanation of the scoring point, acceptable terms (other correct answers), unacceptable terms (partial or otherwise unacceptable responses), and notes (additional comments useful to raters). One of the test developers then scored the summaries using the checklists and used this opportunity to add the acceptable and unacceptable terms in the Guidelines. The examinees' responses were also used to modify the information in the explanation and notes sections of the Guidelines. Then, the second test developer used the draft Guidelines to score several summaries while at the same time adding additional information to the explanations and notes. Following this second scoring the wording of certain checklist points was revised, based on the recommendations of the two scorers. Finally, the project director carefully edited the Guidelines for style and consistency.

After the Guidelines were finalized, the summaries were rescored by one rater using the Checklists and referring to the Guidelines when in doubt. Half of the summaries were then rescored by a second rater in order to gather data on the interrater reliability of the Accuracy score.

3.8 Development of the Expression Rating Scale

Section 2.2.3 gives an introduction to the Expression scale. The Expression rating scale has a history that transcends this particular project. In a previous study by the project director (Stansfield, Scott, & Kenyon, 1992), it was determined that there are two constructs that must be assessed when evaluating a translation. These are Accuracy and Expression. Accuracy refers to the degree to which the translation includes all messages in the source text, while Expression refers to the ability of the translator to convey those messages in writing in the target language.

This finding from the development a Spanish to English translation exam, influenced the approach used to rate summaries on the LSTE-Spanish. However, in a summary translation of a listening text, the accuracy of the messages is most important. The quality of English writing is less important, as long as the message is communicated clearly and accurately so that it can be understood even by those not accustomed to reading the writing of nonnative speakers under these circumstances.[11] Thus, the LSTE-Spanish uses a three-point holistic scale to make gross rather than fine distinctions in the quality of an examinee's writing.

In the development of the LSTE-Minnan, we found a considerable range of writing skills among our examinees. The field test sample seemed to include four groups, those whose writing was often incomprehensible, those whose writing was consistently comprehensible, those whose writing was quite good, and those whose written expression was native. As a result, for the LSTE-Minnan, we expanded the scale to four points: Deficient, Functional, Competent, and Native. It was decided to change the Expression scale in this way after reading the summaries in the process of scoring them for Accuracy. Also, since we were not fully satisfied with the reliability of the Expression rating attained in the LSTE-Spanish,[12] we were willing to experiment with modifications that might improve the reliability of the Expression scale.

3.9 Development of the Expression Scale Self-Instructional Rater Training Materials

Stansfield and Kenyon (1993) have demonstrated that the training of raters need not be carried out by a live trainer. Self-instructional rater training materials can be prepared that

---

[11]For example, in summaries written on the job, run-on sentences and telegraphic style are acceptable.

[12]For the LSTE-Spanish, the interrater reliability for the global Expression rating based on the average rating on the three summaries was .84 (Stansfield, Scott, & Kenyon, 1990, p. 31).

are just as effective as the live trainer. In addition, the self-instructional materials serve as a permanent reference for rater retraining on demand. Because of the success of these materials, it was decided to develop parallel self-instructional rater training materials to accompany the LSTE-Minnan. For the Expression score, the rater training materials were developed as follows.

Once the scale was finalized and the summaries were scored by two raters using it, the ratings were used to develop the self-instructional rater training materials for the Expression scale. This was done as follows:

All summaries were scored twice by two native speakers of English.[13] Since there were three summaries per test form and over 50 examinees took each form, over 300 summaries were scored twice resulting in over 600 Expression ratings. These ratings were then entered into a Paradox database and the database was used to identify summaries on which there was complete agreement across the two ratings. These summaries were then used as a pool from which benchmarks could be drawn. Subsequently, benchmarks were selected for training and testing raters using the self-instructional rater training kit that accompanies the test. Justifications were written for each benchmark selected for inclusion in the rater training kit. They are arranged in the kit by test form and by summary.

---

[13]Barbara Hicks, a language testing specialist with the Evaluation Assistance Center of George Washington University, was the first rater, and the project director was the second rater. Ms. Hicks also wrote the justifications for each of the benchmark summaries included in the training materials.

4.0 Development of Ancillary Data Collection Instrumentation

4.1 Development of Self Assessment Questionnaires

Because no other measures of Minnan were available with which to correlate scores on the LSTE-Minnan, it was decided to develop and use self-assessment questionnaires. A review of the literature on self assessment shows that such measures can be both valid and reliable (Wilson, 1966). In this case, three measures were developed: a Self-Assessment of English Writing Ability (SA-EW), a Listening Comprehension Global Self- Assessment Questionnaire (SA-LC), and a Self-Assessment of Summary Translation Ability (SA-ST).

Each of these self-assessment instruments was drafted by Charles Stansfield, who has worked with the ILR scale for twenty years.[14] The draft was reviewed by project staff and then revised. The revised version was then reviewed by Dr. Pardee Lowe, former Chief of Testing at the CIA Language School, by Marijke Cascallar, Manager of the Foreign Language Education and Measurement Program at the FBI, by Dr. Mats Oskarsson, a professor of language testing at Gotenberg University in Sweden who has over the years published more than anyone on self-assessment of language proficiency, and by Dr. Kenneth Wilson, a senior research scientist at Educational Testing Service who has done research on self-assessment of language proficiency. All made useful comments which were incorporated into the descriptions whenever possible.

4.2 The Self-Assessment of English Writing Ability (SA-EW)

The SA-EW was constructed to imitate the ILR writing scale, but in a format suitable for self-assessment by untrained raters.

---

[14]Because the SA-EW has not been previously validated, it is appropriate to enumerate Stansfield's qualifications to make this adaptation of the ILR skill level descriptions. He has served as an oral proficiency interviewer for the Peace Corps, which uses the ILR scale, and as an oral proficiency interview (OPI) tester trainer for the Peace Corps. He has twice been through ACTFL OPI training (in ESL and in Spanish), and he has conducted several OPI training workshops using the ACTFL scale, which is also an adaptation of the ILR scale. He has also conducted some fifty SOPI (Stansfield, 1989) rater training workshops using the ACTFL scale. He has served, by invitation, as a member of the ILR's Language Testing Committee, and participated in the selection, rating, and justification of the ILR benchmark Reading texts in several languages. He also developed a parallel set of skill level descriptions of translation ability (Stansfield, Scott and Kenyon, 1990) with help from Marijke Cascallar and the ILR Testing Committee.

It was designed to be administered without any accompanying explanation of terms. Therefore, technical jargon for language teachers and references to government work in the ILR skill level descriptions was avoided in constructing each point on the scale.

The format involves a condensed description of only the baseline points on the ILR writing scale. Thus, there is no description of the "plus" levels. This format was chosen because the LSTE-Minnan is essentially a test of listening comprehension in Minnan. English writing ability plays only a minor role in the examinee's performance. In the LSTE-Minnan, the Expression score is considered less important than the Accuracy score. This is reflected in the scoring scale for Expression, which has only four levels, Deficient, Functional, Competent and Native.[15] Because of the amount of precious time that would be required for examinees to read a description of both baseline and plus levels, it was decided not to develop a description of the plus levels. Instead, the plus levels on the SA-EW are represented as being "between" the base levels. With this format, the examinee can read and understand the scale quickly, and can make a fairly accurate self-placement within the scale.[16]

---

[15]These points on the scale are to some degree relatable to levels 1, 2, 3 and 4 on the ILR scale. However, rater judgements that place an examinee on the Expression scale are based solely on the holistic evaluation of short factual summaries. Such summaries constitute a performance test rather that a proficiency test, which is what would truly be required to place a person on the ILR writing scale. Thus, the Expression scale should not be considered a measure of English writing proficiency as reflected in the ILR scale. Nonetheless, because test score users in the Government need to be able to relate the Expression score to the ILR scale, which is a common metric that is used and understood by all government agencies, this information about probable equivalencies on the ILR scale is provided. Since we do not wish to imply that the LSTE Expression score is equivalent to an ILR rating, we use adjectives rather than numbers to refer to the level of the rating.

[16]It should be noted that this format has been previously used by Clark and Swinton (1979) in the development of the Test of Spoken English, and by Oskarsson (1980) in a study of the efficacy of student self-placement within the Council of Europe's unit/credit system. Only Clark and Swinton reported a validity coefficient (.48), which represented the correlation between a single OPI rating and an examinee's self-rating. Oskarsson, however, reported that many students reacted to an initial questionnaire containing only the five ILR-like base levels by placing an X in between the levels and commenting that their skills were "between the two levels." Therefore, Oskarsson

It was decided to have the SA-EW serve as a criterion measure for evaluating the validity of the Expression score. That is, it was assumed that the SA-EW would be an adequately valid measure of English writing skills, and therefore, if the Expression score correlated with it, then that correlation would provide evidence of the validity of the Expression score.[17]

4.3 The Listening Comprehension Global Self-Assessment Questionnaire (SA-LC)

The SA-LC was constructed based on a review of the ILR skill level descriptions for listening and of the ACTFL Proficiency Guidelines for listening.[18]

This particular version of the skill level descriptions for listening has several unique characteristics. The SA-LC was tailored to some degree to the subjects that would participate in the pretesting. Because the subjects would not be government linguists, technical jargon was avoided to the degree possible. In addition, revisions were made in an effort to keep the English employed in the descriptions at a fairly low level (level 2+ or below). In order to reduce the reading load on the examinee, unnecessary repetitions were also deleted. References to memorized utterances and learned material in the lower level descriptions were deleted because they do not apply to native speakers.

At levels four and five, it was decided to use the educated monolingual native speaker of Mandarin as a point of comparison. Because all level 5 speakers of Minnan would have received all or much of their education in Mandarin, it was felt that the

---

modified his scale subsequently to include the between levels option employed here in the SA-EW.

[17]Nonetheless, it should be stated that we did not expect the correlation to be high, since the SA-EW had not previously been validated, and because the 10 point SA-EW scale would be correlated with the Expression scale which has only three points. (A three point scale might not allow adequate differentiation among subjects for it be highly reliable. Thus, even if the SA-EW were a valid and reliable measure, it could not correlate highly with a measure that is lacking in reliability.)

[18]It should be remembered that the ILR listening scale does not specifically identify the overheard conversations tested on the LSTE as a type of listening. Thus, the LSTE focuses on a specific type of listening, while the ILR scale focuses on general listening skills. Nonetheless, the general listening skills associated with the ILR scale appear to be highly relevant to successful execution of the type of listening tasks tested on the LSTE.

comparison with the monolingual Mandarin speaker would convey the high level of listening comprehension skills required to be a level 5 listener.

In Minnan, and even to some extent in Mandarin, the ability to use formal language identifies natives speakers at the higher levels of the scale and differentiates among them. Another feature of the SA-LC was the deemphasis on the ability to comprehend slang, colloquial speech, jokes, puns, and dialectal speech. Native speakers of non-official languages, particularly those without a written form, may acquire the ability to understand slang and colloquial speech long before acquiring the ability to understand professional discussion in the language. Because these aspects of the traditional ILR scale do not apply to native speakers in the same way as they do to nonnatives, they were deemphasized in the SA-LC, but they were not completely eliminated. Their inclusion in all cases where they appear in the ILR skill level descriptions would have presented contradictory and confusion statements to the examinee.

4.4 The Self Assessment of Summary Translation Ability (SA-ST)

The SA-ST was based on a similar instrument that was used in the validation of the LSTE-Spanish. This type of self-assessment was found to correlate highly (.79) with the Total Accuracy score on the LSTE-Spanish. It also correlated highly (.78 for one form and .80 for the other) with the Accuracy score on the Spanish summary writing tasks when the tasks were evaluated by human raters. Thus, it was felt that, since the validity of this self-assessment questionnaire had previously been established, it would be appropriate to employ the SA-ST in the context of the Minnan test as well.

It should be understood that the SA-ST used with the LSTE-Spanish was filled out by FBI linguists who were all experienced at writing summary translations of telephone conversations. Thus, they were all capable of understanding the questionnaire and had ample experience on which to base their self-rating. In the LSTE-Spanish study, the subjects completed the SA-ST questionnaire prior to taking the LSTE. In the case of the LSTE-Minnan, circumstances were clearly different. None of the Minnan pretest examinees had ever written an summary translation of the phone conversation prior to taking the pretest. Thus, it would not have been possible for them to rate themselves on this ability prior to taking the test. As a result, pretest examinees were asked to complete the SA-ST after taking the LSTE-Minnan. At this point, they would have some experience on which to base their self-rating. However, their experience would be limited to only the three summary writing tasks on the test. Thus, it was felt that it would be unlikely that the SA-ST would correlate as highly with the examinee's Total Accuracy score for Minnan as it had for Spanish. Still, it was felt that even a moderate correlation between the SA-ST and the LSTE-Minnan,

would provide evidence of the validity of the latter. Such evidence is useful, since the SA-ST requires that the examinee rate his or her ability to perform the kinds of listening tasks that are often required of law enforcement personnel. Thus, it was felt that a moderate correlation would provide evidence of the relationship between the score on the test and the ability to do the job.

A basic difference between the SA-ST used for Spanish and that used with the Minnan examinees was the addition of a fourth type of conversation to the scale. This was type 4, which involves the ability to understand conversations dealing with scientific, military, or political matters. Although the description of this type of conversation has more to do with topic than with type of speech, it was felt that the addition would be useful in the context of the type of work that actual successful examinees might be asked to perform.

It should be understood that the SA-ST was to address issues of validity within an occupational context. The SA-EW and the SA-LC address the issue of the validity of the Accuracy and Expression scores as indicators of the relevant prerequisite language skills.

5.0 Description of the Field Test Sample

5.1 Background Questionnaire

In addition to the self-assessments discussed in section 4, an examinee Background Questionnaire (BQ) was developed by SLTI (Appendix G). This questionnaire was designed to capture demographic and language background information on the subjects who participated in the field testing of the LSTE-Taiwanese.

The BQ was drafted by staff, based on a similar questionnaire used in the development of the LSTE-Spanish. The draft BQ was reviewed by the FBI before being administered to examinees.

In order to understand factors related to proficiency in Minnan, all participants in the field test program were asked to complete the BQ. Most completed it before their arrival at the test center. Those who did not, completed it after taking the test. The BQ sought five types of information so that we could analyze the background characteristics of the sample that participated in the field testing of the LSTE-Minnan.

I. General information. This includes the age (Q7), name, and institutional affiliation of each examinee (on the cover page of the questionnaire).

II. Language related background. This includes ethnic background and nationality (Q1), birth place of mother (Q2) and father (Q3), and length of time spent in a Minnan-speaking environment (Q9).

III. Language learning and contact. This identifies the individuals from whom Minnan was learned(Q4), and the age at which the learning took place (Q8).

IV. Active language use. These questions identify the interlocutors with whom Minnan is the current means of communication (Q5) and the average amount of time each week when Minnan is used (Q6).

V. Self-assessment of language ability. This requires a self-rating of writing proficiency in English (Q10) and listening comprehension proficiency in Minnan (Q11).

All response options for each question were coded and a database template was created in Paradox. After the data were collected, it was key entered into the database for analysis. A frequency analysis was run using SPSS. The analysis portrayed the raw frequencies, percentage of students responding to each option for each question, and the mean response to each question. The results of the analysis, in terms of the percentage of students responding to each option for each question, are described below.

5.2 Characteristics of Field Test Sample

There were 72 examinees who participated in the field testing. Thus, the total number of responses to the BQ was 72. The analysis of questionnaire data showed that most questions were responded to by 72 individuals. Missing cases, though rare, did occur for some of the questions. Efforts were made during and after the field testing to ensure that every question was answered by all examinees. SLTI staff called several examinees subsequent to field testing to get the responses to questions that were not answered and received the answers over the phone. While some examinees could not be reached, for each question responses were collected on at least 97% of the sample.

5.2.1 Data collection locations

All the participants were from three Universities that have a significant number of Minnan-speaking students: the University of Maryland at College Park (25 participants, representing 34.7% of the sample), the University of California at Davis (19, 26.4%), and the University of California at Berkeley (28, 38.9%). Thus, the sample was fairly evenly drawn from the three institutions, with the largest number coming from UC Berkeley.

5.2.2 Ethnic background and nationality

There are four categories: Chinese-American, which refers to examinees who are US citizens by birth; Chinese from Taiwan; Chinese from the People's Republic of China and US citizens who do not have a Chinese ethnic background. The majority of the examinees (40 subjects representing 55% of the total sample) were from Taiwan. The second largest group is American-born Chinese, occupying a quarter of the sample (18 people, 25%). Only two subjects were from the PRC. These three groups include 88% of the sample. Of the eight people who responded "Other" to this question, four referred to themselves as "Taiwanese American," two were Taiwanese from Japan, and two were Chinese from the Philippines and Canada respectively. In summary, it should be noted that 75% of the sample was born outside the US, while 25% were American born Chinese.

5.2.3 Mother's Place of Birth

Out of the 72 examinees, the mothers of 57 (79%) were born in Taiwan, and 13 (18%) were born in Mainland China. Of the remaining 2 (3%), who chose "Other", the mother of one was born in Hong Kong and the other in the Philippines. Thus, it is clear that the mothers of 4/5 of the sample were from Taiwan. None was born in the US. Thus, in this sample, there were no third generation Minnan speaking subjects.

5.2.4 Father's Place of Birth

Out of the 72 examinees, the fathers of 48 (66.7%) were born in Taiwan, 21 (29.2%) were born in Mainland China. Of the remaining 3 (4.2%), who chose "Other", the father of one was born in the Philippines and two in Japan. Thus, for this sample, 2/3 of the fathers were born in Taiwan. None was born in the US. This is further evidence that in this sample there were no third generation Minnan speaking subjects. It may also be noteworthy that a larger percentage of fathers were born in Mainland China than mothers.

5.2.5 Individuals from whom Minnan was learned

This was determined by the question, "How did you learn Minnan?" Of the five choices for this question (parents, grandparents, relatives, school, other), parents were the most frequent choice (58, representing 81%), followed by grandparents (44, 61%), then relatives (34, 47%), other (19, 26%) and in school (9, 12%). Since the examinee could check all that apply, the percentage from each choice does not add up to 100%. These responses show that 4/5 of the sample learned Minnan from their parents and that very few subjects learned Minnan in school.

5.2.6 Current Minnan interlocutors

Although most subjects learned Minnan from their parents, they seem to speak the language more often with their grandparents, as indicated by the statistics from the analysis. 52 (72.2%) subjects said they currently interact in Minnan with their grandparents, 45 (62.5%) with their mothers, but only 36 (50%) with their fathers, 31 (43.1%) with their friends, and 22 (30.6%) with their sibling(s). Only 4 out of 72 (5.6%) said they speak Minnan with their colleagues. There are 11 cases in which "other" is chosen here. After checking these against the original questionnaire, we found that these include cousins, uncles, and other older relatives associated with the extended family structure. Thus, all of these can actually be considered as "other relatives," which should have been an additional category on the questionnaire. One examinee from the University of California at Berkeley wrote in "teaching friend." Such a response is best considered under the category of "Friends." One of the 11 wrote "work," indicating he must occasionally speak Minnan on the job in the US. It is interesting to note the pattern of disuse of Minnan with successive generations with this group. Nearly 3/4 now uses it with their grandparents, about half with their parents, and about 1/5 with their bothers and sisters.

5.2.7 Amount of time Minnan is spoken in a typical week

There were 68 responses to this question. Thus, there were four non-responses. These were considered different from a value of "0". There were two vague responses (e.g. one half) instead of the number of hours as requested. These were counted as meaning 1/2 hour, although they could have meant "half time". The amount

of time for the valid cases ranges from zero (11 cases, 15% of the sample) to 90 hours (1 case) per week. 54% of the subjects spend between 1 hour (15 cases, 21% of total sample) to 5 hours (8 cases, 12% of total sample) speaking Minnan during an average week. The median for the group was 3.5 hours per week. While Minnan may be used with some regularity, it is clear that it is used a minority of the time by nearly all the subjects.

5.2.8 Age

The subjects may be described as young adults. Of the 72 subjects who participated in the field testing, the youngest was 18 years old and the oldest was 31. The mean age was 20.9 (standard deviation 2.9). Most of the subjects were undergraduate college students between ages 18 and 22 (84%).

5.2.9 Age when Minnan was learned

Most subjects learned Minnan as a child. Actually, 83% learned the language by the time they were 8 years old. Of these 49 (68%) learned it by age 4, and 11 (15%) learned it between 5 to 8. Only 2 of them (3%) said they learned Minnan between 9-14, but as many as 11 (15%) said they learned it at 15 or older.

5.2.10 Time lived in Minnan-speaking area

Almost half the subjects (30, representing 42%) said they have lived in either Taiwan or the Minnan-speaking area of Fujian Province in Mainland China. Fifteen (21%) said they only lived in these countries for 1 year of less. Eight subjects (11%) chose 3-5 years, 4 (6%) chose 2-3 years, and only 3 of them (4%) more than 5 years. Twelve (17%) said they had never been in Taiwan or the Minnan-speaking area in Mainland China. While we did not ask for the examinee's place of birth, the results seem to indicate that about half have lived in China or the PRC. However, the lengths of time indicated by the subjects suggest that most did not spend most of their life there. One must also consider the relatively young age of this largely undergraduate group.

5.3 Results of Self-Assessment Questionnaires

5.3.1 Self Assessment of English Writing Ability

The development of the SA-EW is discussed in section 4.2 of this report. The questionnaire is included in Appendix D. Question 10 on the BQ asked the examinees to record their response to SA-EW. The SE-EW was designed to assess writing ability on an ILR-like scale. Most examinees rated themselves quite high in English writing proficiency. Of the 72 participants in the field testing, the self ratings were as follows.

ILR Level:	1	1+	2	2+	3	3+	4	4+	5
N subjects:		2	2	4	7	14	11	17	15
Percentage:		3%	3%	6%	10%	19%	15%	24%	21%

The results of the analysis show that the median self-perceived English writing proficiency level for the whole group is approximately between levels 3+ and 4. However, these results should be interpreted with caution, as there was no way to accurately verify these results. Indeed, SLTI staff were surprised to see so many subjects rate themselves at levels 4+ and 5.

5.3.2 Self Assessment of Listening Comprehension in Minnan

The development of the SA-LC is discussed in section 4.3 of this report, and is found in Appendix E. Question 11 on the BQ asked the examinees to record their response to the SA-LC. The SA-LC was designed to capture Minnan listening comprehension proficiency on an ILR-like scale. The self ratings were as follows.

ILR Level:	0+	1	1+	2	2+	3	3+	4	4+	5
N subjects:	9	6	13	13	10	6	5	1	7	2
Percentage:	12%	8%	18%	18%	14%	8%	7%	1%	10%	3%

Unlike with self rating of English writing ability, most examinees rated themselves in the lower half of the scale. About half (36, representing 50% of the total sample) rated themselves between 1+ and 2+. 20% rated themselves below level 1+ and only 14% rated themselves above level 3+. The median level was 2. Thus, while it was surprising that the group on the whole felt it could write English so well, it was also surprising that the group did not feel it could comprehend Minnan at a higher level than was indicated, since nearly all might be considered native speakers of Minnan.

5.3.3 Self-Assessment of Summary Translation Ability

The development of the SA-ST is discussed in section 4.4 of this report and is found in Appendix F. After the examinees took the test, they were required to provide a self assessment for their summary translation ability on a 4-level scale: Limited, Functional, Competent and Superior. The number of points on the scale corresponds to the 4-level scale to be used in the Expression rating. The self rating of summary translation ability was obtained following the administration of the LSTE-Minnan because staff anticipated that only at this point would examinees be able to provide such a rating, since prior to doing the summary translation on the test, they had no previous experience performing this type of language task.

Since the summary translation ability may vary according to

the nature of a conversation, each examinee was required to give an assessment for each of the four different types of conversations. Although not included on the BQ, examinee responses to the SA-ST were entered into a data file, and the file was then matched with the data obtained on the BQ. The results for each type of conversation are displayed below.

Type 1.
Conversations in standard Minnan with concrete information (dates, times, locations, amounts, etc.) in a direct manner.

Table 5.3.3.1.
Self-Assessment for Conversation Type 1

Rating	Frequency	Percent
Limited	13	18.1
Functional	14	19.4
Competent	25	34.7
Superior	19	26.4
Missing	1	1.4
Total	72	100.0

These results indicate that most examinees felt they could handle these level 2 conversations fairly well. 61% rated themselves as Competent or Superior in summarizing this type of conversation.

Type 2.
Conversations using a great deal of colloquial language (slang and regionalism) with concrete information (as Type 1) in a fairly direct manner.

Table 5.3.3.2.
Self-Assessment for Conversation Type 2

Rating	Frequency	Percent
Limited	20	27.8
Functional	18	25.
Competent	25	34.7
Superior	8	11.1
Missing	1	1.4
Total	72	100.0

These results indicate that the sample felt it would have some difficulty with slang in Minnan. Few people rated themselves as Superior in dealing with this type of conversation.

Type 3.
Conversations using standard Minnan, possibly with colloquialisms

and making veiled or ambiguous references to shared knowledge (e.g. We'll meet tomorrow at the same place); consequently, very little concrete information may be communicated.

Table 5.3.3.3.
Self-Assessment for Conversation Type 3

Rating	Frequency	Percent
Limited	23	31.9
Functional	21	29.2
Competent	21	29.2
Superior	5	6.9
Missing	1	1.4
Total	72	100.0

The results indicate that the sample rated itself as slightly less capable of providing a summary translation of these conversations than conversations involving extensive use of slang.

Type 4.
Conversations using an educated variety of Minnan to communicate information about political, scientific, or military matters.

Table 5.3.3.4.
Self-Assessment for Conversation Type 4

Rating	Frequency	Percent
Limited	47	65.3
Functional	14	19.4
Competent	9	12.5
Superior	0	0.
Missing	2	2.8
Total	72	100.0

The results indicate than examinees accurately perceived this to be the most difficult type of conversation to handle. The majority rated themselves as having limited ability to summarize such conversations, although about 30% said they could.

5.3.4 Total Summary Translation Ability

SLTI staff decided to come up with an overall rating of summary translation ability. This was the sum of all four self ratings on the instrument. Treating the lowest rating of Limited as having a value of one and the highest rating of Superior as

having a value of four, the range of the scale is from four to 16. The results are presented in the following table:

Table 5.3.3.5.
Self-Assessment for Summary Translation Ability:
Total Numerical Value

Rating	Frequency	Percent
4.00	11	15.3
5.00	6	8.3
6.00	4	5.6
7.00	6	8.3
8.00	4	5.6
9.00	11	15.3
10.00	6	8.3
11.00	9	12.5
12.00	5	6.9
13.00	2	2.8
14.00	3	4.2
15.00	3	4.2
Missing	2	2.8
Total	72	100.0

The results show that perceived summary translation ability was nicely distributed over the sample. It is noteworthy that there was a group of examinees who consistently perceived themselves as having inadequate ability to do summary translation. This is the group with a total rating of 4-6. They constitute about 30% of the sample. Another group perceived themselves as having pretty good ability. This is the group with a total rating of 9-12. They constitute about 43% of the sample. There is another group that rated itself as having high ability to do summary translation. This is the group with a total rating of 13 or higher. This small group constitutes about 11% of the sample. The mean total self rating was 8.6, indicating that the average examinee self rated as Functional across the four types of conversations.

5.4 Summary

The examinee Background Questionnaire and the self assessments proved very useful in understanding the sample that participated in the study. The results indicate that the sample consisted mainly of young adult undergraduates who were born abroad and who learned Minnan at home from their parents and relatives. They continue to use Minnan regularly, although only a few hours per week, and normally with grandparents and other older relatives. The group perceives its English writing ability to be at level 4 or higher,

and its listening ability in Minnan to be mostly limited to informal conversations at level 2 and below. They perceived that they can combine these two proficiencies to produce Competent summary translations in English of low level conversations, but they felt their lack of ability to deal with slang, veiled language, or formal academic/technical terms in Minnan would reduce their perceived effectiveness at summarizing such conversations.

6.0 Field Testing and Revision

6.1 Field Test Administration

The field test version of Forms A and B contained 57 4-option multiple-choice items and three summary translation tasks. The multiple-choice items were based on a total of 12 conversations followed by 3-7 items each. We designed the multiple-choice section of the field test version to include additional items, so that we would be able to delete items that did not perform well from the final version of the test. We also made one conversation and seven items common to each form.

Forms A and B of the LSTE-Minnan were administered at three sites: the University of Maryland, the University of California at Berkeley and the University of California at Davis. These sites were selected because we knew that substantial numbers of Minnan speakers were located there, and because we were able to enlist the cooperation of our colleagues at these universities for cooperation and assistance in recruiting field test examinees. All field test data were gathered between late February and early May, 1996.

The order of administration of the forms was counter-balanced, so that approximately half of the examinees took Form A first and half took Form B first. 52 examinees took Form A and 56 took Form B.

Following the administration, all test booklets, scannable answer sheets, administration instructions, and questionnaires were returned to SLTI via Federal Express. Subsequently, honoraria were sent to each examinee and test administrator. Once the test materials were received by SLTI, the NCS answer sheets were scanned into a database using an NCS Sentry 3,000 scanner.

6.2 Results of Field Testing: Multiple-choice Portions

Following the field test administration, the multiple-choice portions of Form A and Form B were submitted to a classical test analysis and a 1-parameter IRT-based analysis using the BIGSTEPS program developed by Wright and Linacre. Because they are more widely understood, the results of the classical analysis is presented below in Table 6.2.

Table 6.2
Descriptive Statistics for Field Test Versions: Multiple-Choice

Form	Mean	Std. Dev.	Range	Mean Dif.	KR-20
A	41.2	7.3	21-53	.72	.85
B	39	10.2	14-54	.68	.92

Table 6.2 indicates that Form B was slightly more difficult than Form A. The greater difficulty of Form B was confirmed by comments made by examanees to proctors during the pretest administration. The average score on Form B was 39 out of 57 while on Form A it was 41.2. Optimal difficulty on a 4-option multiple-choice test is 62.5% correct. For Form B, the mean represents 68% correct while for form A it represents 72% correct. Thus, both forms were slightly easier than optimal difficulty for this group. The group that took Form B was slightly more varied than the group that took Form A, resulting in a larger standard deviation and a greater range of scores, particularly at the low end of the scale. This greater difficulty and dispersion of scores on Form B resulted in a higher reliability coefficient for Form B. However, the statistics on both forms are good.

Because of the length of the test, and the adequacy of the reliability, we decided to identify the seven least efficient items from each form. The selection of items to be eliminated was based on item discrimination and item difficulty using the classical test analysis. It was also based on a consideration of items with high outfit using an IRT-based analysis called BIGSTEPS. The items that were deleted all had item discrimination indices below .25 and standardized OUTFIT residuals above 2.0.

The seven common items were left intact, for possible use as anchor items in the operational testing program. These anchor items will make it possible to reexamine the item calibrations as the test population changes.

6.3 Results of Field Testing: Summary Writing

Following the collection of the field test data, the summary writing portions of the two forms were scored using the Accuracy Checklists and the Guidelines for Scoring Each Point on the Accuracy Checklists.[19] When scoring the summaries using the checklist, the rater determined if each point was presented in the summary. If a point was presented, then the examinee received credit for a correct answer. If not, the answer was marked incorrect. After all summaries were scored, each examinees answer (correct or incorrect) was entered as either a 1 or a 0 into a database. Subsequently, it was analyzed using both classical and

---

[19]The summaries were scored for Expression by two trained raters using the 4-point holistic scale designed for that purpose. The scoring for Expression is discussed in detail in section 3.8. Because the Expression factor is scored on a holistic scale, no changes in the test were made following field testing. However, subjective information gleaned from a review of field test summaries was used to expand the Expression scale from three points to four.

IRT-based analysis. The results below indicate the outcome of the classical test analysis of the checklist when each point is treated as an item in a test.

Table 6.3
Descriptive Statistics for Field Test Versions: Summary Writing

Form	N Items	Mean	Std. Dev.	Range	Mean Dif.	KR-20
A	54	23.8	9.6	0-39	.44	.92
B	66	17.8	13.8	0-53	.27	.93

As a result of the item analyses of the points in the checklists, 4 inefficient checklist points were deleted from the checklists for Form A and 10 inefficient checklist points were deleted from the checklists for Form B.[20] This resulted in a total of 50 checklist points on Form A and 56 checklist points on Form B.[21]

---

[20]We also subsequently deleted these points and associated information from the *Guidelines for Scoring Each Point on the Accuracy Checklist*.

[21]Because the conversations are not scripted, it is very difficult to create checklists with an equal number of points across the two forms. We eliminated points that did not discriminate, usually because they were so hard that almost no-one included them in the summary, and therefore they did not discriminate different levels of ability. However, we were reluctant to eliminate more points from the summary, since to do so might affect the rater's perception of the validity of the Checklists. That is, if a rater expects to see a point included and doesn't find it included, then the Checklist might appear to be invalid. For example, because the name of the person being called was almost always included in the summaries, it was easy and did not show high discrimination. However, because this information is so essential to a summary, it would be inappropriate to delete it from a checklist.

7.0 Reliability

Reliability refers to degree of consistency in measurement. Fortunately, the reliability of the both forms, sections, and scores on the LSTE-Minnan seems to run from good to excellent. Indeed, the reliability of the LSTE-Minnan equaled or exceeded the reliability of the LSTE-Spanish, which served as its model.

7.1 Reliability of the Accuracy Score

The Accuracy score represents the total amount of information correctly identified. If the examinee takes only the multiple-choice screening test, then the Accuracy score consists of only that subtest. If the examinee takes both the multiple-choice and summary writing sections, then the Accuracy score consists of the score on the summary writing section and the score on the checklists associated with the summaries. We will examine the reliability of the multiple-choice section, the summary writing section, and the combination of the two below.

7.1.1 Reliability of the Multiple-Choice Section

The final version of the multiple-choice section of Forms A and B consists of 50 items. The data on this test are depicted in Table 7.1. The data are based on a reanalysis of the data following the deletion of the less effective items.

Table 7.1
Descriptive Statistics for Final Versions: Multiple-choice

Form	Mean	Std. Dev.	Range	Mean Dif.	KR-20
A	36.4	7.5	16-47	.73	.87
B	34.6	9.7	10-48	.69	.92

Table 7.1 indicates that Form B is slightly more difficult than Form A. The larger standard deviation for Form B suggests that (prior to equating) less competent examinees may have tended to score slight lower and more competent examinees slightly higher on Form B than on Form A. Still the differences are not great.

The mean of Form A represents approximately 73% correct while the mean of Form B represents approximately 69% correct. Thus, for the group as a whole, the tests tended to be slightly easy, since we would expect a mean around 62.5% on a multiple-choice test of optimal difficulty if the sample fully and equally represented the total range of abilities. It may be noted that two lowest scores (10 and 12) were below the chance score of 12.5 on a 50 item test. Thus, when corrected for guessing, these two examinees showed zero ability on the test. A third examinee scored only 15 on Form B.

These three examinees contributed to the lower mean, greater standard deviation, range, and reliability of Form B. Thus, the true differences between the two forms are indeed minor.

It should be noted that while a good range of abilities was found in the sample, the sample contained more high ability students that low ability students as measured by the multiple-choice section of the tests. It should be remembered that the multiple-choice portion was intended to be used as a screen; i.e., to identify candidates who would not do well on the Summary Writing section of the test. Thus, good performance on the multiple-choice section would be a prerequisite to taking the rest of the test. If the total test (MC and Summary) is appropriate for the total sample, then it is not surprising that the multiple-choice section would be slightly easy for the total sample. Thus, the sample does not seem atypical and their high scores on the multiple-choice section are consistent with its intended use.

The internal consistency reliability of the multiple-choice section of both forms of the LSTE is quite good. The reliabilities for the corresponding forms of the LSTE-Spanish were .86 and .88.

The parallel form reliability, the correlation between the score on the two forms, was .87 for a subsample of 29 examinees who took both forms of the LSTE-Minnan.

7.1.2 Accuracy Score: Summaries

The reliability of the Accuracy score on the summaries is depicted in Table 7.2 through a classical test analysis.

Table 7.2
Descriptive Statistics for Final Versions:
Summary Writing-Accuracy

Form	N Items	Mean	Std. Dev.	Range	Mean Dif.	KR-20
A	50	23.6	9.6	0-39	.46	.93
B	56	16.5	13.0	0-49	.29	.96

As can be seen, the data again suggest that Form B is harder than Form A, at least that is the way it turned out for the two samples that took each form on this classical analysis. Both Summary Writing test forms turned were quite difficult for these examinees. Optimal difficulty on this test would be 50% correct, yet the means here represent 47% correct on Form A and 29% correct on Form B. Thus, both tests were harder than is psychometrically optimal for this sample. This was especially true of Form B. Nonetheless, it is noteworthy that two examinees scored very high on Form B. These examinees scored 48 and 49 correctly reported points out of a possible 56, which represents 86% and 88% correct.

The highest score on Form A was 39 correct out of 50, which represents only 78% correct. Thus, for the able candidate, Form B is not unrealistically hard.

Since we have compared this test with the Spanish test in other sections of this report, a comparison here is appropriate too. The mean percent correct for the Accuracy points on the two forms of the Spanish test was 59% and 51%. By comparison, the Spanish test seems to be much easier. However, in absolute terms there may be no difference in difficulty between the Minnan and the Spanish summary writing tests. The differences found here are likely due to the differences in the distribution of summary translation ability in the two language groups. There are many people who have received formal education in both Spanish and English. There is essentially no-one who has received extensive formal schooling in Minnan, one of the two languages tested on the LSTI-Minnan. This clearly would affect the performance of Minnan speakers on a test of summary translation, since they would not recognize many of the words in the source language. So, it is probably the case that there are comparatively few Minnan speakers with high level summary translation ability.[22] For this reason, the LSTE-Minnan should be especially useful to the FBI, since without such a measure, the agency has very little chance of identifying individuals with this ability.

The KR-20 internal consistency reliability coefficients for the Summary Writing section are high (.93 and .96).

The interrater reliability, as calculated on a subsample of half the papers that were scored by the second rater, is extremely high (.99) for both forms. For the LSTE-Spanish, the interrater reliability for the checklists was also very good, ranging from .85 to .93 on the six summaries. However, the almost perfect agreement between raters on the LSTE-Minnan demonstrates that the Scoring Guide for Accuracy makes determining if the answer is right or wrong a highly objective process. This means that only a single rater is needed to score the summaries for Accuracy.

The parallel form reliability for a subsample of 29 examinees who took both forms is also quite satisfactory (.87), although not as high as one might expect given the high internal consistency

---

[22]This conclusion is supported by the results of the self-assesement of listening comprehension in Minnan discussed in section 5.3.2. Only 29% of the examinees self-rated above level 2+, even though nearly all were "native speakers" of the language.

reliability.[23]

7.1.3 Reliability of the Total Accuracy Score

When an examinee passes the screening test and takes the summary writing test, the Total Accuracy score is calculated.

Because the multiple-choice as summary writing tests were treated as separate tests for purposes of statistical analysis, the analyses did not provide an internal consistency reliability coefficient for the Total Accuracy score. Clearly, it would be high, since the number of items on this test is the sum of the multiple-choice items and the checklist points. That total is 100 for Form A and 106 for Form B. Since the reliability of each test is high, the reliability of the combined tests should be very high, probably in the high nineties.

We did correlate the two combined scores for a subsample of 29 examinees who took both forms, and this produced an index of parallel form reliability for the Total Accuracy score. The coefficient was .92.

7.2 Reliability of the Expression Score

Summaries are also scored for Expression. It should be remembered that the Accuracy score is considered the most important of the two for operational needs in the FBI. It is probably appropriate to consider the Expression score as a diagnostic score. That is, the Accuracy score provides the basic information the test user needs, while the Expression score provides information about the examinee on a related trait that is of lesser importance. Nonetheless, operationally the Expression score does have some importance because the examinee's mode of English expression should not hinder the ability of a reader to understand the summary.

Sections 2.2.3. and 3.8. discuss the Expression score. The Expression score consists of the average of the Expression ratings on the three Summary Translations. All of the summaries were

---

[23]In theory, the parallel form reliability should be equivalent to the internal consistency reliability. However, in test development projects where examiness have no stake in their score, it is common for them to loose interest to some extent when taking the test the second time. It is probably the case that that happened here and that the true parallel form reliability is considerably higher than that obtained here. Indeed, when we did a Rasch analysis using BIGSTEPS, the analysis identified two misfitting examinees. When we removed those examinees from the sample the correlation between the two forms increased to .95.

scored by two raters. Then, the ratings on each summary were entered into a Paradox database and Pearson correlational analyses were performed using SPSS. The descriptive statistics and correlation between raters are depicted in Table 7.2.2. for each summary and rater, for the global rating, and for each form.

Table 7.2.2.
Descriptive Statistics for Final Version:
Summary Writing--Expression

Form	Variable	N Cases	Mean	Reliability
A	Sum1,R1	46	2.66	
	Sum1,R2	47	2.68	.85
	Sum2,R1	49	2.35	
	Sum2,R2	49	2.53	.86
	Sum3,R1	49	2.27	
	Sum3,R2	49	2.37	.83
	TotalA,R1	46	2.4	
	TotalA,R2	47	2.52	**.90**
B	Sum1,R1	54	2.69	
	Sum1,R2	53	2.75	.70
	Sum2,R1	53	2.35	
	Sum2,R2	49	2.45	.89
	Sum3,R1	47	2.42	
	Sum3,R2	53	2.42	.81
	TotalB,R1	49	2.48	
	TotalB,R2	45	2.47	**.87**

In the above table, Variable refers to the score obtained when a summary is scored by a rater. Thus, Sum1,R1 refers to the scores on Form A summary 1 assigned by rater 1. N Cases refers to the number of ratings assigned by the rater on that summary. Thus, rater 1 provided 46 scores on summary 1.[24] Mean refers to the mean rating for the variable. Thus, the mean of the scores assigned by

---

[24]The lack of complete data is due to the fact that not all examinees provided an adequate sample for rating their English writing skills. If an examinee did not write a summary or the examinee wrote a very short summary (e.g., "I couldn't understand." or "Lin called Wu."), the rater may not have felt that he or she had an adequate sample with which to make a judgement. In this case, the rater has the option of not assigning a rating. In such cases, it is better not to assign a rating than to assign an incorrect rating. If the lowest rating were assigned it would indicate that the examinee writes poorly in English, when this may not be the case. Rather, it could easily be that the examinee did not comprehend the conversation. Thus, only when a summary of minimal length is produced, is it possible to assign a rating. It is up to the rater to determine if he or she feels confident to provide a rating. If the rater does not provide a rating for one of the summaries, then the global rating for Expression is based on the average of the ratings on the two scored summaries. If a rating is provided on only one summary, then no global rating for Expression is assigned at all.

rater 1 on summary 1 was 2.66. The reliability (in this case inter-rater) is the correlation between the ratings assigned by raters 1 and 2. The reliability coefficients are based only on those cases where both raters assigned ratings.

The total score is the average of the ratings on the summaries. To obtain the total, the ratings were summed and divided by the number of ratings. However, if only one summary was rated, or if no summary was rated, no total score was calculated. The reliability of the total Expression score for a particular form is the correlation between the total Expression scores provided by each rater.

The reliabilities of these short writing samples, usually less than one page in length, is impressive. Only one coefficient, .70 for Summary 1 in Form B, is unimpressive.[25] This coefficient is in fact typical of the interrater reliability one finds on standardized formal writing assessments. For example, the Test of Written English (TWE), of which the Educational Testing Service is justifiably proud, has attained an average interrater reliability of .78 after 10 years of operation (ETS, 1996:10). Five of the six summaries by themselves produced interrater reliabilities that easily exceeded that attained in the TWE program.

The interrater reliability of the global Expression rating is high for both forms (.90 and .87). Such consistency in rating is high enough so that the FBI may feel comfortable relying on only a single rating of Expression. The TWE program, for example, attains reliabilities of this magnitude for the composite rating provided

---

[25]There is no specific reason why this summary produced a lower degree of interrater reliability. Apparently, it was due to the fact that there were two point discrepancies on several papers. These discrepancies might be due to rater error or to characteristics of the individual papers. As of the date of this writing, we have not examined the papers to determine if they show any unique features that might cause raters to disagree as to their quality. In general, we suspect that such disagreements most often occur when a paper is rater a 1 by one rater and a 3 by another. This can happen when a generally competent writer states a particular message in a way that is incomprehensible. In that case, the more severe rater may view the paper as a 1, while the more lenient rater may choose to ignore the incomprehensible message and rate the overall quality of writing instead. It is difficult to provide a general rule as to how this matter should be resolved. However, if the problem continues to occur in the operational testing program, some ground rule for such occasions will have to be developed or perhaps a minor change will have to be made in the wording of the scale. Since the problem only occurred on one of six summaries, we chose not to address it.

by the averaging the ratings of two raters. For the LSTE-Minnan, only a single rating should be necessary to attain this degree of precision in measurement.[26]

---

[26]While the high attained interrater reliabilities are encouraging, it should be noted that they describe only the raters used in this study. Raters in the operational program must be trained using the self-instructional rater training kit for Expression. Individuals vary in the extent to which they can learn to rate reliably. Thus, these reliability coefficients do not apply to any specific future rater. However, these ratings were assigned without the benefit of being trained with the training kit (although the raters subsequently developed the kit). Consequently, we believe that this degree of interrater reliability is generally replicable.

8.0 Validity

According to the Standards for Educational and Psychological Testing (American Psychological Association, 1985), test validity refers to "the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores" (p. 9). Validity is demonstrated by an accumulation of evidence that supports the claim of validity for a particular test. Some of this evidence is based on test performance. Other evidence may be qualitative, in that it deals with the content of the test, or it may be theoretical, in that it deals with a theory about the nature of the trait being measured by the test. In the case of the LSTE-Minnan, the central validity concern is the claim that the test is a measure of the ability to summarize in standard written English the content of a conversation in Minnan.

Traditionally, three types of validity are usually identified according to how the evidence was gathered. These are content validity, criterion-related validity, and construct validity. Construct validity, which "focuses primarily on the test score as a measure of the psychological characteristic of interest" (APA, p. 9), may be understood to subsume the other two types; i.e., content and criterion-related validity are also evidence of the construct validity of a test. We turn first to a discussion of the content validity of the LSTE-Minnan.

8.1 Content Validity

Content validity is evidence that demonstrates the degree to which the sample of items, tasks or questions on a test are representative of the domain of content that could be tested. In the case of the LSTE-Minnan, evidence for its content validity is found in the tasks examinees are asked to perform to demonstrate their ability in listening summary writing. First, the multiple-choice section checks their ability to understand conversations typical of those heard on-the-job. Clearly, without the ability to understand a conversation, there will not be the ability to summarize it. Second, the Summary section checks not only their understanding (the Accuracy score), but also their ability to convey their understanding in written English (the Expression score). In this case, the task directly replicates what is called for on the job. It should be noted that there are two issues here--the accuracy of the information and the acceptability of the English usage in the summary. If the information in the summary is not correct, the summary is of no use to an investigation. On the other hand, if the information is correct but the expression is poor, then the summary could be discredited in a court of law.

Section 3.1., which describes the development of the taped telephone conversations, describes the methodology that was followed to ensure that the conversations on the test simulate actual conversations on the job. The conversations on the test grew out of an analysis of actual conversations taped by the FBI

and the DEA and an interview with an FBI contractor who supervises people doing this kind of work. One of the senior project staff has also done this type of work. Further support for the content validity of the stimuli is provided by the fact that they were based on responses to a telephone interview with FBI language specialists. The analysis of conversational features was then validated by the FBI. Subsequently, a taxonomy of topics and tasks was developed and validated by the FBI. Based on the taxonomy, scenarios were drafted and also validated by the FBI. An FBI special agent who does this kind of work participated in the initial training of actors and the recording of conversations. Finally, the authenticity of the test conversations was validated by FBI language specialists.

8.2 Criterion-related Validity

Criterion-related validity is evidence that "demonstrates that test scores are systematically related to one or more outcome criteria" (APA, p. 11). For example, if there were an extant valid and reliable test of listening summary writing ability, then it would be important to see how scores on the LSTE-Minnan and scores on that test compare. Unfortunately, there is no other test that measures the same construct of listening summary writing ability that could be used as a criterion variable. In fact, no other test of any skill in Minnan exists.

Another direct indicator of criterion-related validity would be to establish a strong relationship between the score on this test and supervisors' ratings of employees' ability to provide good summary translations. For a variety of reasons, this is not possible. First, there was only a small number of FBI linguists working with the Minnan language at the time the study was conducted. Also, none of their supervisors speak Minnan, and thus they are not in a position to provide a valid judgement. Finally, even if such numbers and competent supervisors were available, obtaining such ratings would a difficult and highly sensitive matter.

In the absence of such direct indicators of construct validity, a broad discussion of evidence is important. Such a discussion can be obtained by considering the convergent/divergent nature of the correlations with other measures theoretically related to the construct of interest. In such a discussion, an expected correlation of the test with each variable is analyzed and discussed. Some criteria will be expected to correlate highly with the test whose validity is being examined, while other criteria will be expected to correlate only moderately. Still other criteria might not be expected to correlate at all, or even to correlate negatively. We will make use of the convergent/divergent validity approach here in order to examine fully the construct validity of the LSTE-Minnan.

In an effort to provide evidence for the construct validity of the LSTE-Minnan, data from the self-assessments (discussed in section 5) were correlated with scores on the test, and subscores and sections of the test were correlated with each other. The resulting obtained correlations are depicted and discussed below. For purposes of understanding the variables being discussed, we begin by identifying and defining them below. These variables and correlations are based on the final versions of the test.

SA-LC. Self-Assessment of Minnan listening comprehension on an ILR type scale converted to a numerical value. Maximum score is 10 since 0+ was the lowest point on the scale.
SA-ST. Self-Assessment of summary translation ability total score based on the total of 4 self-ratings using a 4 point scale for each rating. Maximum possible score is 16.
SA-EW. Self-Assessment of English Writing ability on an ILR type scale converted to a numerical value. Maximum score is 10 since 0+ is the lowest point on the scale.
MCA. Score on Form A, Multiple-choice section. Maximum score is 50.
MCB. Score on Form B, Multiple-choice section. Maximum score is 50.
ACCA. Accuracy checklist total, Form A, Rater 1. This is the sum of all points earned for messages conveyed on the three summaries on Form A when they are rated by rater 1. (Rater 2 rated only half the summaries as a check for interrater reliability.)
ACCB. Accuracy checklist total, Form B, Rater 1.
TOTACCA. Accuracy total (MC + checklist) Form A. The Accuracy Total is the sum of correct answers on the multiple-choice and summary translation section, scored by rater 1.
TOTACCB. Accuracy total (MC + checklist) Form B
EXAVALLA. Expression average (composite of ratings by two raters), Form A.
EXAVALLB. Expression average (composite of ratings by two raters), Form B.

8.2.1 Interrelationships between Test Scores

Table 8.2.1. displays the correlations between the multiple-choice section and the checklists. Both are considered to be measures of Accuracy. The numbers in parentheses to the right of the coefficients represent the number of cases (N) that was used to calculate each correlation.

Table 8.2.1.A.
Correlations between MC Accuracy, Checklist Accuracy, and Total Accuracy

	MCA	MCB
ACCA	.85 (49)	.90 (29)
ACCB	.80 (29)	.82 (55)
TOTACCA	.95 (49)	.92 (29)
TOTACCB	.86 (29)	.94 (56)

The above correlations show that there is a high correlation between the multiple-choice section and the checklists for Forms A and B. The strength of this relationship supports the use of the multiple-choice section as a predictor of informational accuracy in the writing of summary translations. Thus, its use as a screening test is validated. (More information on its use as a screen is presented in section 9.5 of this report.)

ASCORE and BSCORE correlate very highly with their corresponding Total Accuracy score TOTACCA (.95) and TOTACCB (.94), of which they form a part. This demonstrates that the MCA and MCB are efficient screening tests for the Total Accuracy score. It also suggest that MCA and MCB could substitute for their corresponding Total Accuracy score; i.e., the scoring of the checklists in order to determine the Total Accuracy score may not even be necessary.

The magnitude of the above relationships also supports combining the multiple-choice section and the summary translation to provide a Total Accuracy score. Further justification for this policy is found by referring to the reliabilities presented in sections 7.1. and 7.1.2. Here, it was noted that the correlation between the two MC forms was .87, and the correlation between the two checklist forms was also .87. The magnitude of these parallel form correlations falls within the range depicted above for cross-section correlations. Thus, it can be observed that correlations across Accuracy sections using different response modalities (MC and checklist) are of about the same magnitude as correlations between sections using the same response modality. Given this data, it is fair to conclude that the two sections tap the same construct with the same efficiency.

Table 8.2.1.B depicts the correlation between the combined number of points earned on the multiple-choice items and the checklists and the composite (average) of the two Expression ratings.

Table 8.2.1.B.
Correlations between Total Accuracy Score (MC+Checklist) and Expression Score

	TOTACCA	TOTACCB	MCA	MCB
EXAVALLA	-.46	-.53	-.41	-.50
	(49)	(29)	(49)	(29)
	P<.001	P<.003	P<.004	P<.005
EXAVALLB	-.06	-.29	-.05	-.30
	(29)	(55)	(29)	(55)
	P<.79	P<.03	P<.78	P<.02

The numbers in parentheses below the coefficients represent the number of cases (N) that was used to calculate each correlation. Two-tailed probability levels (P<.) are indicated below the N. The symbol P< means that the actual probability was slightly less than the number reported here. As indicated earlier, 29 examinees took both forms. The correlations indicate that the Accuracy and Expression exhibit a low to moderate negative correlation. The correlations mean that the greater the proficiency in Minnan (listening), the lower the proficiency in English (writing).

This outcome makes sense, since it is logical that people who have recently arrived in the US would be high in Minnan proficiency and low in English proficiency. Similarly, Minnan speakers who grew up in the US would be high in English proficiency and low in Minnan proficiency. This interpretation represents the extreme cases--people who were born in the US and people who just arrived. If the data were based on this group alone, it is likely that the relationship between the scores would be highly negative. Other examinees who were born abroad but have lived in the US for many years could have good proficiency in both languages, depending on their exposure to both and their efforts to acquire and maintain them. Such factors would produce considerable variability in the proficiencies of this group, with the result that one could expect either no relationship or a low positive relative relationship between language proficiencies in this group. When these different groups of examinees are combined into a single sample as occurred here, it is reasonable to expect a low to moderate negative correlation between the two languages. Thus, the general magnitude, pattern, and direction of the above correlations serve as evidence of the validity of each measure.

8.2.2 Relationships between the LSTE-Minnan Accuracy Scores and the Self-Assessments

Table 8.2.2. below shows the relationships between the multiple-choice sections and the self-assessment measures. The

data permit us to evaluate the validity of the MC sections as a test of summary translation ability. For most examinees (i.e., those that don't pass this test) this will be the only section of the test on which they will be scored.[27] Thus, it is appropriate to evaluate the convergent and divergent validity of this section alone.

Table 8.2.2.A.
Correlations between Self-Assessments and Multiple-Choice Section of LSTE-Minnan

	SA-ST	SA-EW	SA-LC	MCA
SA-EW	-.148 (70) P= .221			
SA-LC	.745 (70) P= .000	-.258 (72) P= .028		
MCA	.745 (44) P= .000	-.276 (46) P= .063	.779 (46) P= .000	
MCB	.764 (53) P= .000	-.117 (53) P= .404	.775 (53) P= .000	.869 (29) P= .000

Table 8.2.2. shows that the MC sections (MCA and MCB) correlate nicely with self-rated listening comprehension skills in Minnan (SA-LC). The correlations are identical when rounded to the nearest hundredth (.78), indicating excellent consistency of measurement for both the MC section and the self-assessment of listening proficiency in Minnan. This fairly high correlation is quite good, since SA-LC is an indirect, rather than a direct measure of listening proficiency. This magnitude of correlation is as good as a test developer could reasonably hope to obtain.

The MC sections also correlate nicely (.75 and .76) with the examinee's mean self-rated summary translation ability (SA-ST) on four types of job-related summary translation tasks involving different types of language and information. The similarity in the

---

[27]The FBI has two options. One is to administer the entire test and score only the MC screening test first. If the examinee passes the MC screen, then the Summary Writing tasks will be scored. The other option is to administer only the MC screening test first and then administer and score the Summary Writing tasks if the examinee passes the screening test.

correlations indicates excellent consistency of measurement for both the MC section and the self-assessment of summary translation ability. Again, this fairly high correlation is quite good, since SA-ST is an indirect, rather than a direct measure of summary translation ability. This magnitude of correlation is as good as a test developer could reasonably hope to obtain. The strength of the relationship with summary translation ability, demonstrates the predictive validity of the MC screening test for predicting performance of a different nature. Although listening and summary translation are two different skills, the LSTE-Minnan MC screening test does an excellent job of predicting summary translation skills.

It is interesting to note that the self-assessments of listening and summary translation skills also correlated highly, indicating that the examinees correctly perceived the strength of the relationship between Minnan listening skills and summary translation of the overheard messages into English.

It is interesting to note the low negative correlations between self-assessed English writing ability and the MC section of the LSTE-Minnan. For neither form is the correlation significantly different from zero. This suggests a zero-to-low negative correlation between these tests and English writing proficiency. The correlations indicate that the MC section of the test is of no utility in predicting English writing proficiency. This prediction based on self-assessments, agrees with the results obtained in the Expression scores. (See Table 8.2.1.B. above.)

Finally, as would be expected based on the data seen thus far, the correlation between the self-assessments of listening and writing is low and negative (-.26). Again, this indicates that all measures (direct and indirect) used in this study functioned consistently in terms of precision of measurement and the construct being measured.

Table 8.2.2.B (next page) below shows the same relationships for the Total Accuracy score. This score is the combined total obtained by adding the MC score and the sum of points earned on the checklists.

The table demonstrates the validity of the Total Accuracy score. The correlations are very similar to those discussed in table 8.2.2.A. above, so there is no need to discuss the interrelationships here. All are as one would hope to find them given everything we know about the sample, the constructs, and the instruments.

Table 8.2.2.B.
Correlations between
TOTAL ACCURACY SCORES (MC + Checklists) and Self-Assessments

	SA-ST	SA-EW	SA-LC	TOTACCA
SA-EW	-.148			
	(70)			
	P= .221			
SA-LC	.745	-.258		
	(70)	(72)		
	P= .000	P= .028		
TOTACCA	.746	-.277	.767	
	(46)	(46)	(46)	
	P= .000	P= .000	P= .069	
TOTACCB	.758	-172	.810	.923
	(53)	(53)	(53)	(29)
	P= .000	P= .218	P= .000	P= .000

It is particularly noteworthy that the correlations are so similar to those involving the MC sections. Indeed, of the seven correlations involving TOTACCA and TOTACCB none of them differs from the MC relationships by more than .05. Two are identical to the nearest hundredth, and two differ by only .01. The average correlation with the self assessments is about .01 higher for Total Accuracy than it is for the MC sections. None of these differences in correlation is significant. This indicates that for the LSTE-Minnan the Total Accuracy score has no greater validity than the MC Accuracy score alone.

This is an important conclusion with major operational implications. It means that one might dispense with the scoring of the summaries using the checklists, unless one wants the slight gain in measurement accuracy that the Total Accuracy score represents. Remember that in section 7.1.1. the KR-20 reliability coefficient for the MC tests was presented as .87 and .92. And, in section 7.1.2. the KR-20 reliability for the Accuracy score obtained on the checklists was .93 and .96 for the two forms. The Total Accuracy score should have a reliability in the mid to high nineties. It would therefore provide for slightly greater precision of measurement at the level of individual scores.

The FBI will have to decide if this slight gain in precision is worth the extra effort of scoring the checklists. Clearly, if this were a large-scale testing program, it would not be worth the additional resources required. However, since this is a very small testing program, the scoring of both sections (MC + summary writing) could be entertained when a person passes the screening test.

Table 8.2.2.C. below present the correlations between the self-assessments and the Accuracy score obtained by summing the points earned on the three checklists that are used to score the three summary translations.

Table 8.2.2.C.
Correlations between the
Self-Assessments and the Accuracy Checklist Scores

	SA-ST	SA-EW	SA-LC	ACCA
ACCA	.696	-.259	.707	
	(44)	(46)	(46)	
	P= .000	P= .083	P= .000	
ACCB	.6874	-.199	.770	.873
	(53)	(53)	(53)	(29)
	P= .000	P= .153	P= .000	P= .000

Again, the correlations are very similar to those obtained with the MC sections alone. Here however, four of the six correlations with the self assessments are slightly lower than were the corresponding correlations obtained with the MC sections. These differences are not statistically significant however.

8.2.3 Summary of Evidence for the validity of the Accuracy scores

The evidence produced in the above section of this report shows that all three measures, the MC section, the summary writing section, and the Total Accuracy score, are valid measures of summary writing ability. In fact, they seem to be about equally valid. Because of this, for purposes of efficiency, the use of only the MC section could be justified.

8.2.4 Relationships between the LSTE-Minnan Expression Score and the Self-Assessments

Table 8.2.4.
Correlations between the LSTE-Minnan Expression Score
and the Self-Assessments

	SA-ST	SA-EW	SA-LC
EXAVALLA	-.5201	.6315	-.5417
	(44)	(46)	(46)
	P= .001	P= .000	P= .000
EXAVALLB	-.1813	.4158	-.2209
	(52)	(52)	(52)
	P= .196	P= .002	P= .115

Table 8.2.4. above shows the correlations between the averaged Expression ratings for each form (3 ratings by 2 raters) and the

examinees' self assessments of Minnan listening, English writing, and summary translation ability. The relationships are moderate and positive for English writing ability; they are low to moderate and negative for Minnan listening and summary translation ability. Again, these directions are of the magnitude and in the directions that one would expect. That is, since Expression is a rating of English writing ability only in the very limited text type of a summary translation, one would not expect a high overall correlation with a more global measure of English writing, such as the self-assessment of English writing on the ILR-like scale that was used in this study. Thus, instead of expecting a high correlation, we expect a moderate correlation, which is what was obtained.

We would expect English writing ability in the restricted context of a summary translation to have some negative relationship with Minnan listening and summary translation ability (both of which we have seen are similar measures of Minnan language proficiency). Indeed, that is what was found here. The moderate correlations were highly significant while the low correlations were nearly significant.

It is noteworthy that the correlations for Form B were of less magnitude that those for Form A. This is because the first summary on Form B produced a lower interrater reliability that any of the others, thereby lowering the reliability of Form B overall. The lower reliability reduced the magnitude of these validity coefficients.

In summary, the Expression score was found to be valid as a measure of English writing ability in the context of summary translation. However, it should be remembered that Expression is best considered as a diagnostic score to be used only with examinees who meet or surpass the pass/fail criterion on the Accuracy score. Examinees who do not meet or surpass this criterion, need not be evaluated for Expression.

9.0 Equating LSTE-Minnan Accuracy Scores

9.1 Equating the Forms

Equating involves an adjustment to test scores when more than one form of the test exists. In the case of the LSTE-Minnan, there are two forms of the test: Form A and Form B, each with two sections. In the multiple-choice section of the test, the number of items on both forms is the same (50). For the Accuracy subscore obtained from the sum of the three summaries scored with the checklists, the number of items is 50 for Form A and 56 for Form B.

We decided to use only the 29 examinees who took both forms for the equating. We would have used all examinees, had we had a larger sample to work with. However, equating based on samples that did not take both forms assumes random assignment of forms and basic equivalency of the ability distribuions as shown by the means and standard deviations for the two distributions. An initial analysis of scores showed that there were significant differences in the means (7 raw score points) and variances of the two distributions for the entire group. However, when we examined the distributions for the common examinees, we found that on the multiple-choice section there was only one point difference in the means for the two forms, and the variances were quite similar as well. On the Accuracy subscore, the difference was two points; however, there was still a large difference in the variances, even for the common examinees.

There are a number of ways in which test forms can be equated. One of the simplist, called mean equating, involves calculating the mean of the distribution of scores on the two forms and then adjusting the raw scores on Form B by adding or subtracting the difference in the mean of the two forms from the Form B score. This gives us the equivalent score on Form A. Thus, if the mean of Form A is one point lower than the mean on Form B (as occured here for the multiple-choice Accuracy score), then one would simply subtract one point from each Form B score to get the Form A score. We felt that this approach was appropriate for the multiple-choice section; however, for a number of reasons we decided not to use mean equating for the Accuracy Checklist scores.

One reason we did not use mean equating for the Accuracy Checklist scores was because Form B contained six more items than Form A. This difference in test length would have produced some anomolous scores.

Another reason why we rejected mean equating for the Accuracy Checklist scores was because mean equating assumes that the difficulty differences in the two forms is constant across the entire score range. In fact, although there was only a two point difference in the means for the Accuracy Checklists, there turned

out to be a seven point difference in the standard deviation on the two forms. Thus, the adjustment of only two points on these section scores would not do justice to these differences.

Another option is linear equating. Linear equating is based on the standardization of the z scores associated with the raw scores obtained on each form. Because the number of items on each form and the standard deviation differs, z scores are used to standardize the raw scores. Then, the z scores are used to equate the raw scores. This means that the forms can be differentially difficult along the scale. Linear equating is more complicated conceptually, but it is more flexible (Kolen & Brennan, 1995). Nonetheless, linear equating suffers from the fact that it is based on a regression coefficient. As a result, score equivalencies near the middle of the distribution are accurate, while equivalent scores near or at the extremes are not. These scores will be considerably closer to the mean than was the obtained score. Linear equating was attempted for the LSTE-Minnan Accuracy Checklist score, but ultimately, after inspecting the degree of regression to the mean, we decided to utilize an IRT approach based on the Rasch measurement model to equate the two LSTE Accuracy checklist forms.

The conversion table for the mean equating of the Form B scores on the multiple-choice section to the Form A scores are found in Appendix K. The conversion table for the Rasch IRT equating of Form B scores on the Summary Checklist to Form A scores are also found in Appendix K. To determine an equivalent total summary accuracy score on Form A based on performance on Form B, the two sections (multiple-choice and Accuracy Checklist) need to be converted seperately and then added together.

The score conversion tables in Appendix K may be used in conjunction with the Raw Score to Summary Accuracy Scale (ILR Equivalent Score) conversion tables that follow in Appendix L. In that case, if interpreting a score on From B one would first procede to Appendix K, where one would find the Form A equivalent score on each section and then total the scores for the Form A Total Score. Then, one would determine the Summary Accuracy Scale (ILR equivalent) score for the Form A score by refering to the score conversion table in Appendix L. Section 9.2, which follows, provides a technical description of how the ILR equivalent scores were derived. The development of holistic descriptions of summary writing ability, which may be useful in interpreting the Summary Accuracy Scale, is discussed in section 9.3.

9.2 Construction of the Summary Accuracy Scale

9.2.1 Overview

In all of the preceding discussion of the LSTE-Minnan, raw

scores have been used. However, one of the goals of the project was to be able to interpret test scores on a descriptive scale. In order for the Government to make decisions on the basis of test scores, compare test scores across forms, and interpret test scores, it is helpful to convert raw scores for Accuracy on the LSTE-Minnan to ILR Equivalent scores. Thus it was necessary to construct a ILR Equivalent score scale for the LSTE-Minnan. This score scale was called the Summary Accuracy Scale. The section that follows describes the rationale for the setting of the ranges of raw Accuracy scores to their corresponding ILR Equivalent scores on the Summary Accuracy Scale.

The first step to achieving this goal entailed the construction of raw score to Summary Accuracy Scale (SAS) score conversion tables for the multiple-choice section scores and the Total Accuracy scores (MC + Summary Accuracy Scores) for each form of the test. These are presented in Appendix L. In this discussion, it must be kept in mind that only **Accuracy** scores are involved. The Expression scores (Deficient, Functional, Competent and Native), which are diagnostic rather than central to the purpose of this test, are **not** converted to an ILR-based scale and are always reported separately from Accuracy scores.

9.2.2 The Selection of the Criterion Variable

Since one of the goals of the project was to provide summary translation ability scores based on a descriptive scale, it was necessary to select an existing ILR-based score that would help anchor LSTE-Minnan scores to the ILR scale. The score was the self-assessment of listening proficiency in Minnan score (SA-LC), which correlated nicely with the three LSTE-Minnan Accuracy measures and was available for all 72 subjects that participated in the field test administration. In addition to the SA-LC score, we had available for 70 subjects a measure that also correlated highly with the three LSTE-Minnan Accuracy scores, namely, the total of the subjects' self-ratings of summary translation ability (SA-ST) on the four types of conversations that government linguists might have to summarize. It should be noted that these two measures also correlated moderately well with each other (.745), showing that the self-assessments were measuring similar although not identical constructs. Plots of the LSTE-Minnan Accuracy measures against both of these variables combined showed that the fit between this combined self-assessment score and the test scores in the critical ranges of ILR 1 to 4 was actually better than that for the self-assessment of listening alone. That is, a small but significant group of subjects in this range performed considerably higher or lower on the LSTE-Minnan than their SA-LC scores alone would predict.

In light of the above, it was decided to use the composite of the SA-LC and the SA-ST as the best indicator of current summary

writing ability. Doing so had the beneficial effect of adjusting the pure listening scores in a way that accommodated them to the additional skills involved in the summary writing task.

To form a composite criterion score for each subject, first all examinees who were missing any LSTE-Minnan Accuracy or self-assessment scores were eliminated from the data set. This left 70 subjects for consideration. Second, to ensure equal weighting in the composite score, the two self-assessment scores were transformed into standardized z scores using the mean and standard deviation for each distribution. Thus, they were linearly transformed to have a mean of 0 and a standard deviation of 1. The third step was to add the two standardized scores together. Finally, this total self-assessment score composite was scaled through a linear transformation to correspond back to the ILR scale. This transformation used two anchor points. The first was the highest possible raw score on the two measures (10 on the SA-LC and 16 on the SA-ST, which was equal to a z score of 4.4784 on the composite). This was assigned a level 5 on the ILR-based Summary Accuracy Scale. The second anchor was the "minimally competent" score (5 on the SA-LC and 8 on the SA-ST[28], which was equal to a z score of .0111 on the composite). This was assigned to a level 2+(2.6) on the Summary Accuracy scale.[29]

The formula for a linear transformation is

$$\text{scale score} = A \times \text{raw score} + B$$

where A is the slope (i.e., scaled score 2 - scaled score 1/raw score 2 - raw score 1) and B is the intercept (i.e., scaled score 2 - A x raw score 2). By substituting the equivalencies given above, the following equation was derived for converting the composite scores to the ILR-based scale score:

---

[28]The SA-LC went from 0+ to 5. Thus, counting from the bottom, 2+ was the fifth point on the scale. Therefore, a 2+ was assigned a numerical value of 5 in the database. Similarly, a rating of Limited was the second point on the SA-ST scale. It was assigned a value of 2. Because there were four items on the SA-ST, the numerical value of 8 corresponds to an overall rating of Limited.

[29]Composite z scores are based on the following data using the anchor points.

	SA-LC		SA-ST		Composite z	ILR
	10		16			
Raw z	2.2043	+	2.2738	=	4.4785	ILR 5
	5		8			
Raw z	.2004	+	-.1893	=	.0111	ILR 2.6

ILR-based scale score = (.5373 x composite z score) + 2.5939

In this way each examinee received an ILR-based Summary Accuracy Scale score for accuracy in summary writing ability.

To see how this score fit the test data better, we can compare the relationship between the individual parts. This is presented in Table 9.1 below. The numbers in parentheses represent the N for each correlation.

Table 9.1
Correlations of Self-Assessments of Listening Comprehension (SA-LC) and Summary Translation (SA-ST) and Composite (SA-LC+SA-ST) Score Converted to the Summary Accuracy Scale (SAS) with the LSTE-Minnan Accuracy Raw Scores

	SA-LC	SA-ST	SAS
MCA	.78 (46)	.75 (44)	.82 (44)
MCB	.78 (53)	.76 (53)	.83 (53)
ACCA	.71 (46)	.70 (44)	.75 (44)
ACCB	.77 (53)	.69 (53)	.79 (53)
TOTACCA	.77 (46)	.75 (44)	.81 (44)
TOTACCB	.81 (53)	.76 (53)	.85 (53)

Note: N (in parentheses) = all examinees with complete data

Table 9.1 shows that using the composite score for SAS consistently gave a better fit to the test data than the SA-LC alone. Thus basing the SAS score on the composite of the two self-assessments provides a better foundation for building a score conversion table.

9.2.3 Outliers Detected and Removed

The next step was to determine exactly which examinees and scores would be actually included in the scaling and subsequent equating. In terms of inclusion, it was noted that for some examinees we had scores on both forms, while for others we had scores on only one. We decided that each examinee would be given equal voice by being included only once. We also decided to use the examinee's Form A score when both were available. For examinees who took only Form B, we decided to use the Form B raw score equated to the Form A raw score. For each section of the test and for the total score, these raw scores were regressed against the SAS scores.

The preliminary examination of the raw score data revealed

that there were some highly influential cases. This was particulary true for cases at the extremes; i.e., persons who gave themselves very high self-assessment scores on the accuracy checklist portion. (This is one reason why the SAS scores were developed rather than using a straight SA-LC score. However, it remained to be seen whether there were still any outliers in the set whose test performance behavior can not be explained by using the SAS score.) Inclusion of these overly influential outliers in the data set to convert LSTE-Minnan scores into SAS scores might jeopardize the usefulness of the results for score interpretation and decision making.

To detect these influential outliers, Cook's Distance statistic, which is available in SPSS-PC, was used. Those cases with a Cook's D value above .05 were removed. Of the 70 subjects on which there were data on all measures, seven were identified as outliers in the MC sections in the first run. These subjects were eliminated from the data for the final regression run. Five outliers were deleted from the checklist Accuracy distribution. For the Total Accuracy score, Cook's D identified seven cases as overly influential. These were likewise deleted from the data for the final regression runs.

9.2.4 Effects of Removing Outliers

The elimination of these cases improved the correlation between the three Accuracy measures and the Summary Accuracy Scale as depicted in Table 9.2 below. Therefore, for each subtest and for the Total Accuracy score, these subjects were deleted from the data set before proceeding to develop the conversion tables for each score.

Table 9.2.
Correlation of Equated Accuracy Scores with Summary Accuracy Scale Prior to and after Deletion of Outlying Examinees

	Prior	Afterwards
MC	.82	.86
ACC	.80	.84
TOTACC	.84	.89

9.2.5 Development of Raw Score to Summary Accuracy Score Conversion Tables

From the correlations in Table 9.2, three regression equations were derived. These equations were then used to predict ILR-based SAS scores from multiple-choice section scores, from Total Accuracy Checklist scores, and from Total Accuracy Scores. These three conversion tables are available to potential test score users in Appendix L.

The following comments about the score points in these conversion tables should be noted:

1. For the multiple-choice sections, a score of 13 or below can be achieved by chance. Thus, there is no SAS equivalent for those scores.

2. One can convert effectively from both the multiple-choice section and the Accuracy Checklist score to the SAS. However, the most accurate measurement is on the basis of the Total Accuracy score, for two main reasons. First, the total score, which is a composite of the two section scores, contains more variance and a wider spread of scores than the multiple-choice section score alone. Second, the total score correlated slightly higher (.03) with the SAS score than did the multiple-choice section alone. (See Table 9.2.)

3. As is true whenever regression equations are used, the most accurate conversions will be around the mean of the scales. The means obtained in this study were near the cut-off Summary Accuracy Scale score of 2.6. Thus, we can be especially confident about the accuracy of the conversion tables at and near the cut score. The SAS conversion table is less accurate at the extreme ends of the range of LSTE-Minnan scores.

9.3 Further Analyses of the SAS Scores.

The scatterplots in Appendix M may provide further understanding of the relationship between the ILR-based Summary Accuracy Scale and scores on the test. The points on the regression line show the predicted SAS score for future LSTE-Minnan examinees for each test score; the scattered points indicate the actual observations from this sample. (Note that these plots include all observations; influential observations - outliers - have not been deleted.) From each scatterplot, we can make the following observations.

1. The regression line has a fairly good fit at most points in the distribution. However, high scoring examinees had very inconsistent self-assessments. That is, there was a wide distribution of self-assessment scores for high-scoring examines. Some gave themselves fairly high self-assessments (tending towards 4 to 4+), while other high scoring examinees gave themselves self-assessments around 2+ and 3.

2. The effect of dispersion of the self-assessment scores at the upper end of the ability distribution is that the prediction of the ILR-based Summary Accuracy Scale score is less precise in this region. As a result of this dispersion, the regression line itself never reaches the ILR score of 5, even for a perfect score on any

section of the test. Thus, in the conversion tables in Appendix L, no scores convert to a score of 5 (or even 4+ in some cases).

3. The accuracy of the regression around the important cut-off scores (see Section 9.2.5 and Appendix L) is higher than at the extremes since the cut-off scores are closer to the mean of the ability distributions. For the suggested cut-off score of 30 on the MC section, it may be noted from the plots that only one examinee scoring below 30 on the multiple-choice Section gave herself a self-assessment rating above 2.6. However, 18 individuals scoring above 30 gave themselves a self-assessment rating below 2.6. Thus, this conservative cut-off score of 30 ensures that individuals who do view themselves as having high proficiency in Minnan will be given the complete test. It also ensures that persons who feel their skills are weak but score above the cut score, will be given the chance to further demonstrate their ability on the performance-based measure.

4. The fact that the SAS scale does not reach the highest levels may not be relevant, since there may not be anyone who in fact has level 5 skills in Minnan. Few people have been educated in Minnan, since it is not an official language. Thus, in theory there are no educated native speakers of the language.

9.4 The Final Accuracy Rating

In order to give an interpretive description to the 0-5 Summary Accuracy Scale scores, a Final Accuracy Rating scale was developed during the development of the LSTE-Spanish, which served as the prototype for the LSTE-Minnan. The Final Accuracy Rating scale is based on six descriptions of summary writing ability: No Ability, Severely Deficient, Deficient, Functional, Competent and Superior. (These categories are similar to the self-assessment of summary translation ability categories, which ranged from Deficient to Superior.) The scale was developed based on two sources of input. The first was the discussion of errors in accuracy (misinterpretations, omissions, and additions) in the FBI/CAL translation skill level descriptions developed for a previous project and published in Stansfield, Scott, and Kenyon (1992). The second was the range of performance of examinees in terms of the quantity of accurately reported details in the summaries they wrote. The six descriptions on the Final Accuracy Rating scale thus represent holistic performance descriptions that were written in reference to the translation skill level descriptions and to natural performance groupings within the sample tested for the LSTE-Spanish. However, we believe that the these holistic descriptions apply to the LSTE-Minnan as well. For that reason, the Final Accuracy Rating descriptions are included in Appendix J.

The cutting point for the rating of No Ability was developed considering the chance score on the multiple-choice section. There were fifty multiple-choice items; thus the chance score on this

section is 13. This level of performance or lower represents no ability. The cut score for the remaining descriptions is the point at which the corresponding Summary Accuracy Scale score exceeds .60. Although the use of .60 is essentially arbitrary, this value was selected because it corresponds to the lower-bound value of a plus level when ILR levels are converted to numerical values. Thus, the remaining cutting scores for converting to the Final Accuracy Rating scale are the raw scores that are equivalent to a SAS score of 1.60, 2.60, 3.60 and 4.60. The range for the Incompetent category goes from the cutting score for No Ability to 1.59; for the Deficient category it goes from 1.60 to 2.59; for the Functional category from 2.60 to 3.59; for the Competent category from 3.60 to 4.59; and finally for the Superior category it goes from 4.60 to 5.0.[30]

While the correlations between the Accuracy and the multiple-choice sections were high (.85 and .82 on Forms A and B respectively), only the summaries represent performance samples from which a performance description can be extracted. Still, given the high correlation between the two sections, and the similarity of the listening stimuli and the type of information tested by the multiple-choice section (main topic, key points, and supprting details), it is probably appropriate to use the Final Accuracy Rating performance descriptions to interpret performance on any of the accuracy measures; i.e., the multiple-choice section, the checklists, or the multiple-choice section and the checklists combined (Total Accuracy).

It may be useful at some point in the future for the FBI to perform a cross-validation analysis of the Final Accuracy Rating descriptions in Appendix J for the LSTE-Minnan. In other words, an analysis could be carried out of the performance of examinees in each category on the FAR scale in terms of the average number of points they identified in their summaries. These actual mean performance levels could then be compared with the FAR scale descriptions.

9.5 Using the Multiple-Choice Section as a "Screen"

The multiple-choice section of the LSTE-Minnan may be used to screen out individuals for whom the Summary section of the test is inappropriate; that is, examinees would not be likely to have a Total Accuracy score at a 2.6 or above on the summary accuracy scale (Functional or above on the FAR). In this case, the most serious error to make in using the multiple-choice section score is to make a decision to exclude someone from taking the Summary

---

[30]Note that a 4.6 occurs only for a perfect score on the accuracy checklist. This is because there were too few cases at the extremes of this scale to measure Superior level skills with precision.

section who would pass it, rather than to give the Summary section to someone who may not ultimately receive a FAR of Functional. To determine the cut-off score on the multiple-choice section, we need to first determine the raw score on the multiple-choice section that corresponds to a Functional score (2.6 on the SAS). Once this is found, we then need to determine the lowest possible raw score one could get on the multiple-choice section while, given measurement error, still having a statistical possibility of scoring at that cut-off score level.

The raw score on the multiple-choice section that most closely corresponds to a passing score of 2.6 is 36 on Form A (which is a score of 37 on Form B). Given the reliability of the two tests at .87 and .92 respectively and the variances of the equivalent samples of 29 common examinees (7.69 for Form A and 7.93 for Form B), the standard error of measurement (SEM) for Form A is 2.77 and for Form B it is 2.24. Thus, the 95% confidence interval around the passing score would then be:

Form A	36	-	2 x 2.77	to	36	+	2 x 2.77	= 30.46 to 41.54
Form B	37	-	2 x 2.24	to	37	+	2 x 2.24	= 32.52 to 41.48

This means that an examinee scoring 30 or below on Form A of the multiple-choice section or 32 or below on Form B has less than a 2.5% probability of having a "true" raw score of 36 or 37, respectively, on each form, which corresponds to a 2.6 on the SAS. Because the Form B raw score is converted to the Form A raw score, 30 is the cut-off scores for the multiple-choice section. Examinees who score below this level on the multiple-choice section of the LSTE-Minnan either need not take the Summary section, or if they already have, that section need not be scored.

Using this cut-off score will still leave in many examinees who may not ultimately achieve a Final Accuracy Rating above Functional; however, the chance of excluding a candidate who might achieve a Functional is slim.

As a final comment, it is obvious that scores on the multiple-choice section cannot predict Expression scores. That is, a candidate may achieve a passing score on the multiple-choice section (and on the Final Accuracy Rating), yet ultimately not pass the LSTE-Minnan on the basis of a Deficient Expression score. The multiple-choice section of the LSTE-Minnan is **not** intended to screen out such candidates.

References

American Council on the Teaching of Foreign Languages (ACTFL). (1986). *ACTFL Proficiency Guidelines.* Yonkers, NY: ACTFL.

Clark, J.L.D. & Swinton, S.S. (1979). *An exploration of speaking proficiency measures in the TOEFL context.* TOEFL Research Report 4. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1966). *TOEFL Test of Written English guide.* Princeton, NJ: Educational Testing Service.

Interagency Language Roundtable (ILR). 1985. Interagency Language Roundtable language skill level descriptions. In Duran, R.P, Canale, M., Penfield J., Stansfield, C.W., & Liskin-Gasparro, J., *TOEFL from a communicative viewpoint on language proficiency.* TOEFL Reseach Report 17. Princeton, NJ: Educational Testing Service.

Kolen, M.J. & Brennan, R.L. (1995). Test Equating Methods and Practices. New York: Springer.

Oskarsson, M. (1980). *Approaches to self-assessment in foreign language learning.* Oxford: Pergamon Press.

Scott, M.L., Stansfield, C.W. & Kenyon, D.M. (1966). Examining validity in a performance test: the Listening Summary Translation Exam (LSTE)--Spanish version. *Language Testing, 13(1),* 83-110.

Stansfield, C.W. (1989). *Simulated Oral Proficiency Interviews.* ERIC Digest. Washington, DC: ERIC Clearinghouse for Languages and Linguistics.

Stansfield, C.W., Scott, M.L., & Kenyon, D.M. (1990). *Listening Summary Translation Exam (LSTE) - Spanish.* Final Project Report. Washington, DC: Center for Applied Linguistics. ERIC Document Reproduction Service, ED 323 786.

Stansfield, C.W., Scott, M.L., & Kenyon, D.M. (1992). The measurement of translation ability. *Modern Language Journal, 76,* 455-67.

Wilson, K.M. (1996). *Validity of global self-ratings of ESL speaking proficiency based on an FSI/ILR-referenced scale: An empirical assessment.* Draft final report. Princeton, NJ: Educational Testing Service.

Abbreviations and Abbreviation Equivalencies

In order to standardize the interpretation of the LSTE technical reports, this report has used abbreviations identical to or similar to those used in the LSTE-Spanish report. However, these standardized abbreviations were not employed at the time the database created and the statistical analysis was run. Because the statistical analyses for this study were turned in to the FBI, the list of abbreviations used in this report and their corresponding abbreviations used in the statistical analyses are enumerated here below. The abbreviation on the left refers to the abbreviation used in this report. It is identical or very similar to the abbreviation used in the LSTE-Spanish report. The abbreviation on the right is used in the statistical analysis printouts accompanying the deliverables.

SA-LC = SALISNUM. Self-Assessment of Minnan listening comprehension on an ILR type scale converted to a numerical value. Maximum score is 10 since 0+ was the lowest point on the scale.
SA-ST = SASTTOT. Self-Assessment of summary translation ability total score based on the total of 4 self-ratings using a 4 point scale for each rating. Maximum possible score is 16.
SA-EW = SAWRTNUM. Self-Assessment of English Writing ability on an ILR type scale converted to a numerical value. Maximum score is 10 since 0+ is the lowest point on the scale.
MCA = ASCORE. Score on Form A, multiple-choice section. Maximum score is 50.
MCB = BSCORE. Score on Form B, multiple-choice section. Maximum score is 50.
ACCA = ACTOTAJJ. Accuracy checklist total, Form A, Rater 1 (Jing-Jing Liu). This is the sum of all points earned for messages conveyed on the three summaries on Form A when they are rated by rater 1.
ACCB = ACTOTBJJ. Accuracy checklist total, Form B, Rater 1 (Jing-Jing Liu)
TOTA = ACTOTALA. Accuracy total (MC + checklist) Form A. The Accuracy Total is the sum of correct answers on the multiple-choice and summary translation section.
TOTB = ACTOTALB. Accuracy total (MC + checklist) Form B
EXPA = EXAVALLA. Expression average (composite of ratings by two raters), Form A.
EXPB = EXAVALLB. Expression average (composite of ratings by two raters), Form B.
SAS = SELFILR. The Summary Accuracy Scale constructed for this study based on a scaling of the sum of the subject's SA-LC and SA-ST. SAS is reported on a 0-5 ILR-based scale.

BEST COPY AVAILABLE

APPENDIX A

MULTIPLE-CHOICE SECTION TEST BOOKLET
(Selected Pages)

Examinee Name (LAST,First):_____Test Booklet #___

SOUTHERN FUKIENESE (Taiwanese) Listening/Summary Translation Exam

Form A

April 5, 1997

Part A: Multiple-Choice Questions

Developed by

Second Language Testing, Inc. (SLTI)

With Funding from the

Center for the Advancement of Language Learning

PART A: Multiple-choice ITEMS

Introduction:

In this section, you will hear a series of conversations of varying lengths. For each conversation, there are several multiple-choice items. Both the question and the four choices are printed in your test booklet. Before you hear the conversation, you will be given an opportunity to briefly scan the questions. This will show you what type of information to listen for. Before each conversation, read ONLY the questions; you will not have enough time to read the possible answers.

After listening to the conversation, read the questions again. Then read the four possible answers for each question carefully and select the best one based on what you have heard in the conversation. Locate the number of the question on your answer sheet and fill in the space that corresponds to the letter of the answer you have chosen.

Example:

1. What is the mobile phone number?

   A. 311-9014
   B. 311-9024
   C. 322-9014
   D. 322-9024

2. When are they going to meet?

   A. In seven or eight hours
   B. The next day
   C. This week
   D. Next week

Explanation:

In the conversation, Lim told Teng that his mobil phone number is 322-9014. Therefore, the correct answer to question 1 is choice (C).

Lim told Teng in the conversation that this week will be better for a meeting, but since Teng cannot make it, next week will be a good time too. Thus the best answer to the second question is "Next week", which is choice (D).

Remember you may NOT take notes or write in your test booklet during this part of the listening section. Mark your answers by darkening the spaces on your answer sheet.
CONVERSATION 1 (C#9)

APPENDIX B

SUMMARY WRITING SECTION TEST BOOKLET
(Selected Pages)

Examinee Name(LAST, First): _____Date: ___/___/___

SOUTHERN FUKIENESE (Taiwanese) Listening/Summary Translation Exam

Form A

April 5, 1997

PART B: SUMMARY WRITING TASKS

Introduction:

In this part of the test, the conversation will be heard twice. As you listen to the conversation, write down important information in the space marked "NOTES". Check the accuracy of the information as you listen the second time. Then, within the time limit, write a summary of the conversation in the space marked "SUMMARY". Try to provide as much detail as possible, without making your summary a word-for-word translation of the conversation.

Important information includes the general purpose of the conversation and supporting details, such as main points, names, dates, times, places, amounts, and other concrete information. The conversations vary in the amount of concrete information they contain. If a conversation deals with an abstract topic, make sure you identify the general topic and include supporting information.

After each conversation, you will have a limited amount of time to write a summary of the conversation in as much detail as possible. The amount of time you are given to write a summary will depend on the length of the conversation, which will range from 1 to 4 minutes. You will be informed of how much time you will be given to complete each summary.

Sample notes and summaries:

NOTES

Teng. San Francisco; Lim. L.A. Lim waiting for Teng. several weeks; when's Teng coming? business needs to be done; ASAP; Teng should come next week. if not this week; Teng: how long it takes to drive to L.A. how to contact Lim; Lim: eight hours to L.A. call 322-9016 when he arrives.

SUMMARY

TENG in San Francisco calls LIM in Los Angeles, who has been waiting for his call for several weeks. LIM wants to know when TENG will be coming because their business needs to be completed as soon as possible. They agree that TENG should come next week if he cannot make it this week. TENG wants to know how long it takes to drive to L.A. and how to contact LIM upon arrival. LIM tells him that it takes seven to eight hours and asks him to call him at 322-9014 when he arrives.

Explanation: Notice the important information that was included in the summary:

The names of both parties in the conversation are written in CAPITAL LETTERS. Other information includes **who calls who, the main purpose of the call, and factual information**. This summary includes the location of each party, the fact that LIM has been waiting for the call for several weeks, when DENG should come, the time required to get there, and the telephone number where DENG should contact LIM upon arrival.

Notice that the summary has been written in full sentences and in paragraph form, rather than as a list. Remember, your score will depend on the accuracy and completeness of the information in your summary, and on how well it is written in English. Therefore, if you finish your summary before the time limit, you should check your English spelling and expression.

APPENDIX C

EXPRESSION SCORING GUIDE

Scoring Guide for Expression: LSTE-Taiwanese

Deficient	At this level, the writer is not consistently able to communicate with the reader. Errors are numerous and some interfere with communication. The summary may include errors of grammar, spelling, punctuation, and sentence fragments. The writer may rely on simple grammatical and syntactic structures to communicate and even these may contain errors. The Deficient level writer can be understood only by someone accustomed to writing produced by English language learners.
Functional	At this level, the writer is able to communicate meaning in a generally successful way. Errors, although frequent, normally do not interfere with communication. Errors typically occur in syntactic structures, (verb form, subject-verb agreement, choice of tense, coordination of tenses, use of articles and prepositions) and in spelling.
Competent	At this level, the writer is able to communicate clearly with few errors. The summary can be read and understood without effort. The writer is able to produce complex sentences and to convey the messages unambiguously, although the writing may be awkward on occasion. The summary may contain organizational features such as transition words (e.g., but, however) or a topic sentence. Some idiomatic expressions may be used.
Native	At this level, the writer is able to communicate as smoothly and as effortlessly as a well educated native writer of English operating under the same test situation constraints. The writer demonstrates a range of vocabulary and syntax, along with appropriate idiomatic usage and register. Errors are rare and if made, are the type of errors that would be made by a native writer. Such errors never interfere with communication or disturb the reader.

APPENDIX D

SELF-ASSESSMENT OF ENGLISH WRITING ABILITY

Name_____ Institution_____

Self-Assessment of English Writing Ability

Please rate your ability to write in English. Below are ten descriptions of different levels of writing ability, ordered from low to high. Please circle the number of the description that best corresponds to your level. Then, on the background questionnaire, in the space following question 10, please write the number you have circled.

0+ My English writing ability is less than 1.

1. **I can only write short notes or messages, post cards, and simple letters.** My vocabulary, grammar and spelling are inadequate; therefore, I must write in simple sentences. My English writing is limited to practical necessities.

1+ My ability is between 1 and 2.

2. **I can handle in writing everyday correspondence (letters), take notes, write summaries, factual descriptions, and factual narratives.** My command of English vocabulary and grammar, although limited, is adequate to handle such tasks in a minimally adequate way. I make frequent errors in punctuation and grammar.

2+ My ability is between 2 and 3.

3. **I can express myself in writing effectively on most practical, social, and professional topics.** I can write social and business letters, reports, summaries, short library research papers on current events or particular areas of interest with reasonable ease. My control of punctuation, grammar, spelling, and vocabulary is adequate to convey my message. I can use compound and complex sentences and I can present ideas clearly.

3+ My ability is between 3 and 4.

4. **I can express myself precisely and accurately in writing on social issues or my educational or professional needs, using a variety of prose styles.** I can tailor my writing to my audience and express subtleties and nuances. I can express any of my experiences and ideas. I have a broad vocabulary and I rarely make grammatical mistakes.

4+ My ability is between 4 and 5.

5. **I can generate formal and informal correspondence, such as official reports and documents, with the competence of a published writer.** I can write for professional purposes, such as legal, technical, educational or literary purposes. My writing is clear, explicit, informative, and, if necessary, original. I can employ a wide variety of prose styles or rhetorical approaches, as required by the purpose and audience. I can edit my own writing to the point where there are no grammatical, spelling, or punctuation errors, and I can edit the work of others too.

APPENDIX E

SELF-ASSESSMENT OF LISTENING COMPREHENSION

Name_____ Institution_____

Self-Assessment of Listening Comprehension in Minnan

Please rate your ability to understand spoken Minnan. Below are ten descriptions of different levels of ability, ordered from low to high. Please circle the number of the description that best corresponds to your level. Then, on the background questionnaire, in the space under question 11, please write the number you have circled.

0+. **I can only understand short phrases, utterances and sentences, expecially when the context strongly supports understanding and the speech is clearly audible.** I can comprehend *words and phrases from simple questions, statements, common commands,* and expressions of courtesy. I can understand some references to basic personal information or the immediate physical setting. When spoken to in Minnan, I require repetition, rephrasing, and a slowed rate of speech.

1. **I can only understand utterances about basic everyday needs, such as meals, lodging, transportation, time, simple instructions and directions.** I can understand *simple questions, answers and statements* in simple face-to-face conversations in a standard dialect if they are spoken more slowly and clearly than normal, with frequent repetition or rewording.

1+. **I can normally understand short conversations about survival and travel needs and some longer stretches of speech dealing with current, past, and future events.** I can understand *simple descriptions of places and precise instructions*. I sometimes have to ask for utterances to be repeated. I can understand common verb and question forms and word order patterns in the language, but my comprehension fails with complex language forms and patterns. I can understand simple telephone conversations and clear careful speech, such as simple radio announcements.

2. **I can understand face-to-face speech in standard Minnan spoken at a normal pace, with some repetition and rewording, even when spoken by a native speaker who is not used to speaking to people with limited proficiency in Minnan.** I can understand speech dealing with *everyday topics,* common personal and family news, well-known current events, and routine work matters. I can understand descriptions of different places and *narrations about current, past, and future events*. I can follow the essential points of an elementary discussion on work-related topics in my field. I can understand the facts, such as reported in news broadcasts, but I do not grasp inferred meanings or implications expressed through more complex language.

2+. **I can understand speech in standard Minnan dealing with routine social situations and most work-related conversations.** I can understand *some professional discussions* on concrete topics related to my fields of interest. However, I can not sustain comprehension of longer discourse which is linguistically complex or deals with abstract topics. I have some ability to infer meanings not directly stated.

3. **I can understand the essentials of all talk in standard Minnan including discussions within my field of interest or specialization.** I can follow accurately the essentials of conversations between educated native speakers of Minnan, reasonably clear telephone calls, *radio broadcasts, news stories, oral reports, some oral technical reports, and public presentations on nontechnical subjects.* I can infer meanings that are not directly stated.

3+. **I can understand most of the content as well as the intent of professional discussions in Minnan, discussions on general topics, and social conversation.** I can follow accurately the details of conversations between educated native speakers of Minnan, including *conversations on technical subjects.* I can understand native Minnan speakers talking quickly, using regional dialect or slang. I can infer meanings and normally grasp implications. I can understand some subtleties and nuances with social or cultural meanings within the language.

4. **I can understand all forms and styles of speech pertinent to my social and professional needs.** This includes speech involving *extensive and precise vocabulary*, subtleties and nuances in standard dialects of Minnan, and *technical discussion on professional topics within the range of my knowledge.* I can understand language tailored to different audiences and purposes, including persuasion, representation, counseling, and negotiating. I can readily infer meanings and implications. I can easily understand all social conversations, radio broadcasts, and phone calls. I may experience some difficulty understanding speech heard under unfavorable conditions, such as through a poor quality loudspeaker or radio, or in a noisy room.

4+. **I can almost always understand educated and academic speech, abstract professional or academic discussions, regional dialects or slang, and speech heard under unfavorable acoustic conditions.** My comprehension of Minnan is *almost always equivalent to the Mandarin comprehension of a well-educated monolingual native-speaker of Mandarin.*

5. **I can fully understand educated and academic speech, abstract and professional discussions, regional dialects, highly colloquial speech, jokes and puns, and speech in noisy places or heard under unfavorable acoustic conditions.** My comprehension of Minnan is *fully equivalent to the Mandarin comprehension of a well-educated monolingual native-speaker of Mandarin.*

APPENDIX F

SELF-ASSESSMENT OF SUMMARY TRANSLATION ABILITY

Name_____ Institution_____

Self-Assessment of Summary Translation Ability

Now that you have experience in writing summary translations in English of conversations you have heard in Minnan, please provide a realistic evaluation of your ability to do such tasks. Your responses will be used to assess the effectiveness of this exam. Please estimate your ability to summarize four types of conversations using the scale below.

Limited	I can correctly report in my summary about half of the key points of information conveyed.
Functional	I can correctly report the topic of the conversation; however, my summary may contain misinterpretations or omission of several key points.
Competent	I can correctly report the topic and most key and supporting points.
Superior	I can correctly report all key points and a lot of supporting details, including nuances of tone and emotion.

Now, based on all the conversations you have heard, please evaluate candidly your ability to summarize the different types of conversations described below by circling the appropriate label for each.

Type 1 In Type 1 conversations, speakers generally use standard Minnan to communicate concrete information (dates, times, locations, amounts, etc.) in a direct manner.

Limited **Functional** **Competent** **Superior**

Type 2 In Type 2 conversations, speakers use a great deal of colloquial language (slang and regionalisms) to communicate concrete information (as above) in a fairly direct manner.

Limited **Functional** **Competent** **Superior**

Type 3 In Type 3 conversations, speakers use standard Minnan, possibly with colloquialisms, and make veiled or ambiguous references to shared knowledge (for example, "We'll meet tomorrow at the same place at the same time"); consequently, very little concrete information may be communicated.

Limited **Functional** **Competent** **Superior**

Type 4 In Type 4 conversations, speakers use an educated variety of Minnan to communicate information about political, scientific, or military matters.

Limited **Functional** **Competent** **Superior**

APPENDIX G

TEST-TAKER BACKGROUND QUESTIONNAIRE

Test-taker Background Questionnaire

We would appreciate your answer to the following brief questions concerning your Minnan language background. Your answers will help us to identify those factors that are related to proficiency in Minnan.

Name: _____

Institution: _____

Email Address: _____

Phone Number: (_____) _____

Fax Number: (_____) _____

Postal address (for mailing payment): _____

_____

1. Please specify your ethnic background and nationality.

___ a. Chinese-American (US citizen by birth)
___ b. Chinese from Taiwan
___ c. Chinese from PRC
___ d. US citizen, not Chinese
___ e. Other, please specify _____

2. Where was your mother born?

___ a. Taiwan
___ b. Mainland
___ c. USA

3. Where was your father born?

___ a. Taiwan
___ b. Mainland
___ c. USA

4. How did you learn Minnan? (Indicate all that apply.)

___ a. From parents (one or both)
___ b. From grandparents (one or both)
___ c. From relatives
___ d. From school
___ e. From others (please specify) _____

5. With whom do you interact in Minnan? (Indicate all that apply.)

___ a. Father
___ b. Mother
___ c. Grandparents (one or both)
___ d. Brothers and Sisters
___ e. Friends
___ f. Colleagues
___ g. Others (Please specify)

6. Please estimate the amount of time you speak Minnan in a typical week.

_____hours

7. Please indicate your age.

_____years

8. At what age did you learn Minnan?

___ a. 0-4 years
___ b. 5-8 years
___ c. 9-14 years
___ d. 15 years or older

9. How long have you lived in Taiwan or in the Minnan-speaking area of Fujian province?

___ a. I have never been there.
___ b. 1 year or less
___ c. 1-2 years
___ d. 2-3 years
___ e. 3-5 years
___ f. more than 5 years

10. Enter here the number you circled (0+ - 5) on the Self Assessment of English Writing Ability. _____

11. Enter here the number you circled (0+ - 5) on the Self Assessment of Listening Comprehension in Minnan. _____

APPENDIX H

LSTE-MINNAN EXAM FEEDBACK QUESTIONNAIRE
Pre-Field Testing

Questionnaire for Minnan-Speaking FBI Linguists

Background:

This questionnaire is to be used with a cassette tape which contains 19 short simulated telephone conversations in Minnan and two longer ones. The conversations are being considered for use on a Listening Summary Translation Exam in Minnan that is being developed by Second Language Testing, Inc. for FBI Headquarters. Your responses to these questions will help us assess whether the conversations are suitable for the test.

Instructions:

Insert the tape in the tape recorder; then, push the Play button to begin the tape. After listening to each conversation, stop the recorder and circle the letter of the most appropriate response to each question. There are seven questions for each conversation.

Now, begin listening to the first conversation (C#34) and respond to the questions that follow it.

Conversation C#34.

A. In comparison with most Minnan conversations you listen to on the job, how similar are the **linguistic features** (voices, manner of speech, naturalness) of this conversation?

1. Very similar
2. Similar
3. Dissimilar
4. Very dissimilar

B. In comparison with most Minnan conversations you listen to on the job, how similar is the **topic** of this conversation?

1. Very similar
2. Similar
3. Dissimilar
4. Very dissimilar

C. In comparison with most conversations you listen to on the job, how **difficult to understand** is this conversation?

1. Easier than most
2. About average
3. More difficult than most

D. In comparison with most conversations you listen to on the job, how **clear** is the tape recording of this conversation?

1. Clearer than most
2. About average
3. Less clear than most

E. In comparison with most conversations you listen to on the job, how fast is the **rate of speech** for this conversation?

1. Faster than most.
2. About average.
3. Slower than most.

F. Do you think this conversation is **appropriate for inclusion** on a test of listening proficiency in Minnan?

1. Very appropriate
2. Appropriate
3. Inappropriate
4. Very inappropriate

G. Optional. Please write any **other comments** you wish to make about this conversation in the space below.

APPENDIX I

INSTRUCTIONS TO FIELD TEST ADMINISTRATORS

Test Administration Instructions

Listening Summary Translation Exam
Southern Fukienese (Taiwanese) Version

Developed by
Second Language Testing, Inc.
N. Bethesda, MD 10852
Phone 310-231-6046
FAX 310-231-6046
email: Charlie@cal.org

With Funding from the
Center for the Advancement of Language Learning (CALL)

NOTE TO TEST ADMINISTRATOR

This manual describes important information about procedures that must be followed before, during and after the administration of the translation exams. Uniform procedures are essential for the translation exams to yield reliable test results. The scores of all examinees taking the test at different sites around the nation will be comparable only if all test administrators follow the same procedures and give exactly the same instructions. It is necessary, therefore, that you become read the entire manual before administering the exams, that you become thoroughly familiar with these instructions, and that you follow the instructions without exception when administering the exams.

GENERAL INFORMATION

Test Security

It is extremely important that the translation exams be safeguarded and administered under secure conditions at each test site. In order to ensure test security, it is essential that you adhere to the following conditions:

1. Keep all test materials either in your immediate physical possession or in a locked cabinet or other secure area under your control.

2. Do not copy, or allow others to copy, any portion of the test booklets or tape, or make any notes or transcription on the test booklets or tape content.

3. Allow only those particular individuals who are to be tested to see the test materials, and only at the time of test administraiton and under the specific procedures described in this manual.

4. Should any irregularities occur, report them on the Test Administrator Report Form included in this manual.

PRIOR TO THE TESTING DATE

Assembling Test Materials

Assemble as many test booklets and answer sheets as will be needed for this particular upcoming administration, and an extra copy of each. You should also have on hand two sharpened no. 2 pencils (with erasers) for each examinee. Listed below are the materials needed for the Listening Summary Translation Exam:

1) Multiple-Choice Section test booklets
2) Summary Section test booklets
3) Answer sheets
4) Sharpened No. 2 pencils
5) Two copies of the tape for the form to be administered.
6) A high quality cassette playback unit (unless the test will be administered in a language laboratory).
7) A stapler
8) A small pencil sharpener, if one is not located in the testing room.

Arranging for a Testing Site

Unless the test will be administered in a language laboratory, locate a testing site that is comfortable and free from distraction. The listening exam requires a quiet room with good acoustics throughout and a high quality cassette exam

playback unit. The testing room should be large enough so that examinees can be seated with three feet of space in all directions between all examinees.

Language lab. If the test will be administered in a language laboratory, leave at least one empty booth on each side of each examinee in order to prevent cheating or distraction.

Equipment

Check the playback equipment to make sure that it is functioning properly. Adjust the volume control so that everybody in the room can hear the recording clearly. If the playback unit has a tone control, it should be set to the middle ("flat response") position or adjusted somewhat toward the treble. It should not be turned toward the bass position Make sure that the tape is completely rewound after making these adjustments. Be sure to have two copies of the test tape on hand in case of breakage or malfunction.

Language lab. Check each booth that will be used by an examinee. Make sure that the equipment functions properly. This includes the headset plugged into the booth. Set up the test tape in the master console and play it out to the booth. Verify that it can be heard through the headset in each booth where an examinee will be tested. If not, take appropriate action and correct the problem.
Note: The use of individual cassette tape playback equipment that may be installed in each booth is discouraged. Under such an approach to administering the test, some examinees may try to stop and rewind the cassette. This constitutes cheating, and for this reason, the central console should be used to play the test to each examinee.

Prohibited Materials

Examinees may not use dictionaries during the Multiple Choice section; however, they may use either a bilingual or English language dictionary during the Summary writing section.

ON THE DAY OF THE TEST

Administering the Test

Follow the procedures below when administering the test. All bolded instructions should be read VERBATIM. Do not depart from these directions unless noted otherwise.

1. As soon as each examinee arrives ask if he or she brought with them their completed questionnaires. These include the Examinee Background Questionnaire, the Self-Assessment of English Writing Ability, and the Self-Assessment of Listening Comprehension Ability in Minnan. If they did not yet complete

the questionnaires or if they did not bring them with them, give them each missing questionnaire and have them fill out the questionnaires now if there is time, or at the end of the test if you are ready to start. Advise them of this now and tell them they will not be paid for their participation in the testing program unless they complete the entire test and turn in all questionnaires.

2. When you are ready to begin administering the test, inform the examinees:

The Listening Summary Translation Exam in Taiwanese lasts approximately two hours. All of the instructions for filling out the answer sheet are given on the test tape. There will be an opportunity to ask questions before the actual test begins.

3. Distribute the test booklet, answer sheets and pencils.

4. Give the following instructions:

Please do not open your test booklet. In this section of the exam, you may mark your answer in the test booklet and then transfer them to the answer sheet. You must use a no. 2 pencil for marking your answers.

5. Begin playing side A of the tape.

6. Make sure the test tape form (A or B) corresponds to the test booklet form.

7. Walk around the room to make sure that everyone is fouled out the answer sheet correctly as they take the test.

8. After the examinees have answered Item 57, inform them:

This is the end of the Multiple Choice section. Please stop working now. Now look over your answer sheet carefully. Be sure all the marks you made are dark and heavy. Make sure your name and date are written in English on the front cover of the test booklet. Now insert your answer sheet in your test booklet and close your test booklet.

9. Now turn over the tape to Side B. Rewind briefly to the start position.

10. Immediately collect the Multiple Choice test booklets and answer sheets.

11. Distribute the Summary Writing test booklets and ask examinees to write their name on them.

12. Begin playing Side B of the tape. (All instructions for the

BEST COPY AVAILABLE

Summary Writing section are given on the tape.)

13. At the end of the test, inform the examinees:

Please stop working now. Close your test booklet.

14. Now distribute the Self Assessment of Summary Translation Ability. Now inform the examinees:

Please take out your questionnaires. Complete the Examinee Background Questionnaire now, if you did not do so before.

Examine the rating you assigned yourself on the Self-Assessment of Listening Comprehension in Minnan. Read the description or the level you assigned yourself, then read the descriptions immediately above it and immediately below it. Decide if you wish to change your rating, based on your experience in taking this test. If you change your rating, be sure to indicate this in the appropriate space on your Examinee Background Questionnaire.

Allow at least one minute for the above activity. Then say the following.

Now enter your name and date on the Self Assessment of Summary Translation Ability and read the directions for this questionnaire. Assign yourself a rating on the four types of conversation at the bottom of the page.

Allow at least two minutes for the above activity. Then say the following.

Now make sure that your name is on all the questionnaires and on the Summary Writing test booklet. Put all the pages of your Summary Writing test booklet in order. I will collect the test booklet and all the background questionnaires from you now. After I have collected them from everyone, I will make one additional announcement. If you need additional time to complete the questionnaires, you may continue filling them out after the others have left.

15. Go to each examinee with the stapler. Staple the all the pages of the Summary Writing test booklet and make sure the examinee's name is on it. Collect all four questionnaires: the Examinee Background Questionnaire, the Self-Assessment of English Writing Ability, the Self Assessment of Listening Comprehension Ability in Minnan, and the Self Assessment of Summary Translation Ability. Make sure that each questionnaire contains the examinee's name.

16. Now inform examinees.

You will receive your payment in the mail within 10 days. Within 90 days, we will send you an analysis of your

performance on the test. Thank you very much for your participation in the development of this test of Taiwanese. You may leave now or after you have completed all questionnaires and turned in all test materials.

17. Now complete the Test Administrator Report Form.

18. If you have any questions or experience any problems you may contact SLTI at the address on the cover of the test booklet.

19. Return the answer sheets, the test booklets, the questionnaires, and the Test Administrator Report Form in a Federal Express box using the Federal Express tracking label you were provided. You may take the box to a Federal Express branch office or to any office of Mailboxes USA. Charge the shipment to account number 183-763-298. Send via the standard overnight rate or the two day ecomomy rate. Be sure to collect and save your pink receipt for the materials. Mail all materials to

Dr. Weiping Wu
Center for Applied Linguistics
1118 22nd St. NW
Washington, DC 20037
ph. 202-429-9292
email: jingjing@cal.org

20. Please feel free to send us a memo with any comments or observations about the test or these Test Administration Instructions. Comments about the test could reflect your own observations or those of examinees.

21. As soon as all materials are received, you will be sent a check for your participation in this project and for any expenses you incurred.

APPENDIX J

INTERPRETATION OF 0-5 FINAL ACCURACY RATING SCALE

INTERPRETATION OF FINAL ACCURACY RATING

NO ABILITY	No response or fails to identify overall topic accurately. Typically provides no substantial information beyond names of speakers.
SEVERELY DEFICIENT	Often fails to identify topic accurately. Contains frequent misinterpretations, omissions, and/or misleading additions. Usually less than a fourth of the key points of information are correctly reported.
DEFICIENT	May not represent topic accurately. Contains many misinterpretations, omissions, and/or misleading additions. About half of the key points of information may be correctly reported.
FUNCTIONAL	Normally identifies topic accurately; however, contains misinterpretation, omission, and/or misleading addition of several key points of information. May contain a number of supporting details.
COMPETENT	Accurately reports almost all key points of information and many supporting details; no misleading additions.
SUPERIOR	Accurately reports all or almost all key points of information and supporting details.

APPENDIX K

SCORE CONVERSION TABLES

FORM B RAW SCORE TO FORM A RAW SCORE FOR MULTIPLE CHOICE SECTIONS AND ACCURACY CHECKLISTS

Mean Equating from Form B scores to Form A scores
On the Multiple Choice Section: FINAL TABLE

Form B Score	Equivalent on Form A	Form B Score	Equivalent on Form A
0	0	26	25
1	0	27	26
2	1	28	27
3	2	29	28
4	3	30	29
5	4	31	30
6	5	32	31
7	6	33	32
8	7	34	33
9	8	35	34
10	9	36	35
11	10	37	36
12	11	38	37
13	12	39	38
14	13	40	39
15	14	41	40
16	15	42	41
17	16	43	42
18	17	44	43
19	18	45	44
20	19	46	45
21	20	47	46
22	21	48	47
23	22	49	48
24	23	50	49
25	24		

Rasch Equating from Form B Scores to Form A Scores
On the Accuracy Checklist: FINAL TABLE

Form B Score	Equivalent on Form A	Form B Score	Equivalent on Form A
0	0	29	30
1	3	30	31
2	5	31	31
3	7	32	32
4	9	33	33
5	11	34	33
6	12	35	34
7	14	36	34
8	15	37	35
9	16	38	35
10	17	39	36
11	18	40	37
12	19	41	37
13	20	42	38
14	21	43	38
15	22	44	39
16	22	45	40
17	23	46	40
18	24	47	41
19	24	48	42
20	25	49	43
21	26	50	43
22	26	51	44
23	27	52	45
24	27	53	46
25	28	54	47
26	29	55	49
27	29	56	50
28	30		

APPENDIX L

SCORE CONVERSION TABLE: SUMMARY ACCURACY SCALE (ILR-BASED)

Note: These conversion tables take the LSTE-Minnan scores and convert them to an equivalent score on the 0-5 ILR scale that is commonly used in the US Government. Interpretation of this scale, in the context of summary translation, is assisted by referring to Appendix J. In Appendix J, the rating of No Ability is equivalent to a 0 on the Summary Accuracy Scale, while a rating of Superior is equivalent to a 5 on the Summary Accuracy Scale.

Conversion to ILR Scores: FINAL TABLE
Multiple Choice Portion
page 1/2

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
1	***	0
2	***	0
3	***	0
4	***	0
5	***	0
6	***	0
7	***	0
8	***	0
9	***	0
10	***	0
11	***	0
12	***	0
13	***	0
14	0.64	0+
15	0.73	0+
16	0.82	0+
17	0.91	0+
18	1.00	1
19	1.10	1
20	1.19	1
21	1.28	1
22	1.37	1
23	1.46	1
24	1.55	1
25	1.64	1+

*** = chance scores

Conversion to ILR Scores: FINAL TABLE
Multiple Choice Portion
page 2/2

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
26	1.73	1+
27	1.82	1+
28	1.91	1+
29	2.00	2
30	2.09	2
31	2.18	2
32	2.28	2
33	2.37	2
34	2.46	2
35	2.55	2
36	2.64	2+
37	2.73	2+
38	2.82	2+
39	2.91	2+
40	3.00	3
41	3.09	3
42	3.18	3
43	3.27	3
44	3.37	3
45	3.46	3
46	3.55	3
47	3.64	3+
48	3.73	3+
49	3.82	3+
50	3.91	3+

Note: Regression Standard Error = .438

Conversion to ILR Scores: FINAL TABLE
Accuracy Checklist Portion
page 1/2

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
1	0.99	0
2	1.07	1
3	1.14	1
4	1.22	1
5	1.29	1
6	1.37	1
7	1.44	1
8	1.52	1
9	1.59	1
10	1.67	1+
11	1.74	1+
12	1.82	1+
13	1.89	1+
14	1.97	1+
15	2.04	2
16	2.12	2
17	2.19	2
18	2.27	2
19	2.34	2
20	2.42	2
21	2.49	2
22	2.57	2
23	2.64	2+
24	2.72	2+
25	2.79	2+

Conversion to ILR Scores: FINAL TABLE
Accuracy Checklist Portion
page 2/2

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
26	2.87	2+
27	2.94	2+
28	3.02	3
29	3.09	3
30	3.17	3
31	3.24	3
32	3.32	3
33	3.39	3
34	3.47	3
35	3.54	3
36	3.62	3+
37	3.69	3+
38	3.77	3+
39	3.84	3+
40	3.92	3+
41	3.99	3+
42	4.07	4
43	4.14	4
44	4.22	4
45	4.29	4
46	4.37	4
47	4.44	4
48	4.52	4
49	4.59	4
50	4.67	4+

Note: Regression Standard Error = .480

Conversion to ILR Scores: FINAL TABLE
Total Scores (Multiple Choice + Accuracy Checklist)
page 1/4

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
1	0.08	0
2	0.13	0
3	0.17	0
4	0.21	0
5	0.26	0
6	0.30	0
7	0.35	0
8	0.39	0
9	0.44	0
10	0.48	0
11	0.52	0
12	0.57	0
13	0.61	0+
14	0.66	0+
15	0.70	0+
16	0.75	0+
17	0.79	0+
18	0.84	0+
19	0.88	0+
20	0.92	0+
21	0.97	0+
22	1.01	1
23	1.06	1
24	1.10	1
25	1.15	1

Conversion to ILR Scores: FINAL TABLE
Total Scores (Multiple Choice + Accuracy Checklist)
page 2/4

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
26	1.19	1
27	1.23	1
28	1.28	1
29	1.32	1
30	1.37	1
31	1.41	1
32	1.46	1
33	1.50	1
34	1.55	1
35	1.59	1
36	1.63	1+
37	1.68	1+
38	1.72	1+
39	1.77	1+
40	1.81	1+
41	1.86	1+
42	1.90	1+
43	1.94	1+
44	1.99	1+
45	2.03	2
46	2.08	2
47	2.12	2
48	2.17	2
49	2.21	2
50	2.26	2

Conversion to ILR Scores: FINAL TABLE
Total Scores (Multiple Choice + Accuracy Checklist)
page 3/4

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
51	2.30	2
52	2.34	2
53	2.39	2
54	2.43	2
55	2.48	2
56	2.52	2
57	2.57	2
58	2.61	2+
59	2.65	2+
60	2.70	2+
61	2.74	2+
62	2.79	2+
63	2.83	2+
64	2.88	2+
65	2.92	2+
66	2.97	2+
67	3.01	3
68	3.05	3
69	3.10	3
70	3.14	3
71	3.19	3
72	3.23	3
73	3.28	3
74	3.32	3
75	3.36	3

Conversion to ILR Scores: FINAL TABLE
Total Scores (Multiple Choice + Accuracy Checklist)
page 4/4

Score on Form A	Predicted Self-Assessment ILR Score	ILR Equivalent
76	3.41	3
77	3.45	3
78	3.50	3
79	3.54	3
80	3.59	3
81	3.63	3+
82	3.68	3+
83	3.72	3+
84	3.76	3+
85	3.81	3+
86	3.85	3+
87	3.90	3+
88	3.94	3+
89	3.99	3+
90	4.03	4
91	4.07	4
92	4.12	4
93	4.16	4
94	4.21	4
95	4.25	4
96	4.30	4
97	4.34	4
98	4.39	4
99	4.43	4
100	4.47	4

Note: Regression Standard Error = .405

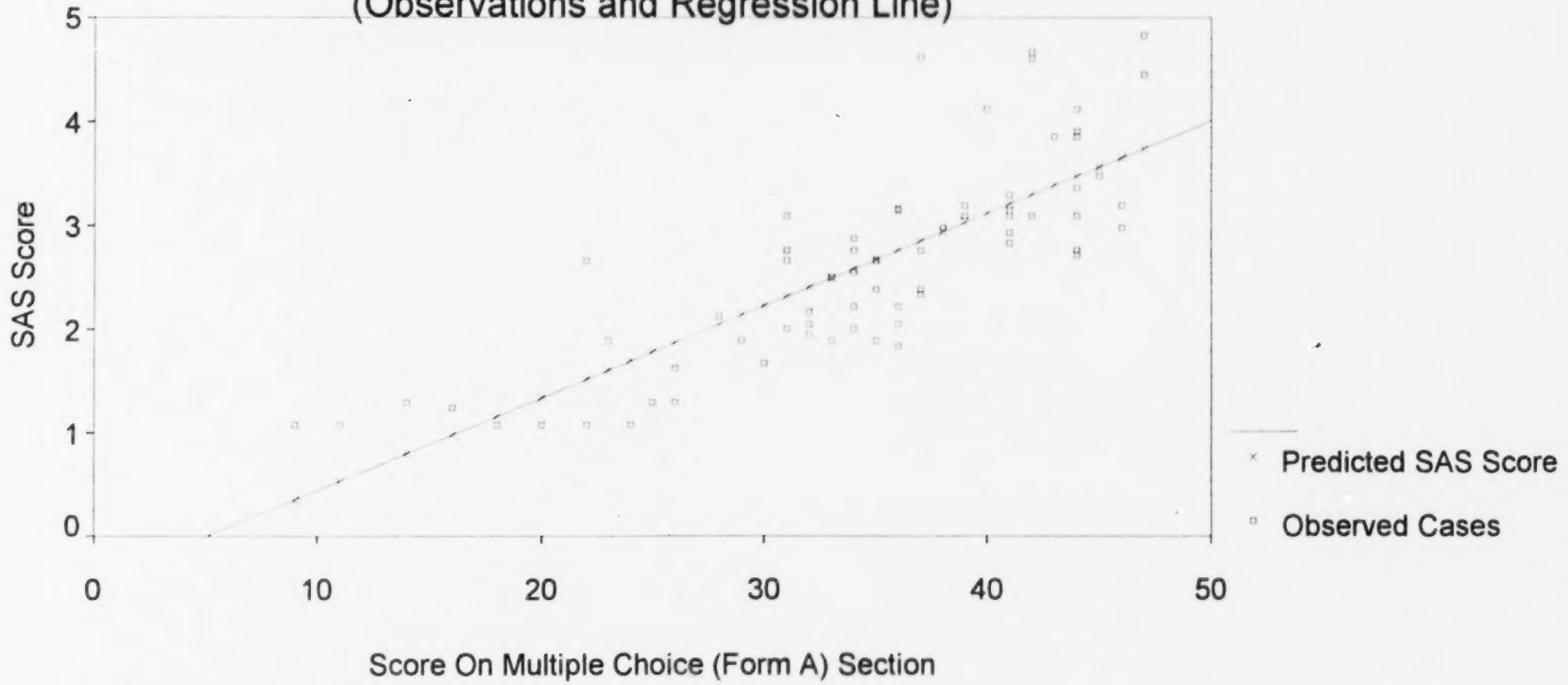
APPENDIX M

SCATTERPLOTS DEPICTING PREDICTED AND OBSERVED SAS SCORES FROM TEST SCORES

ILR-Based SAS Score

Predicted from MC (Form A) Scores

(Observations and Regression Line)

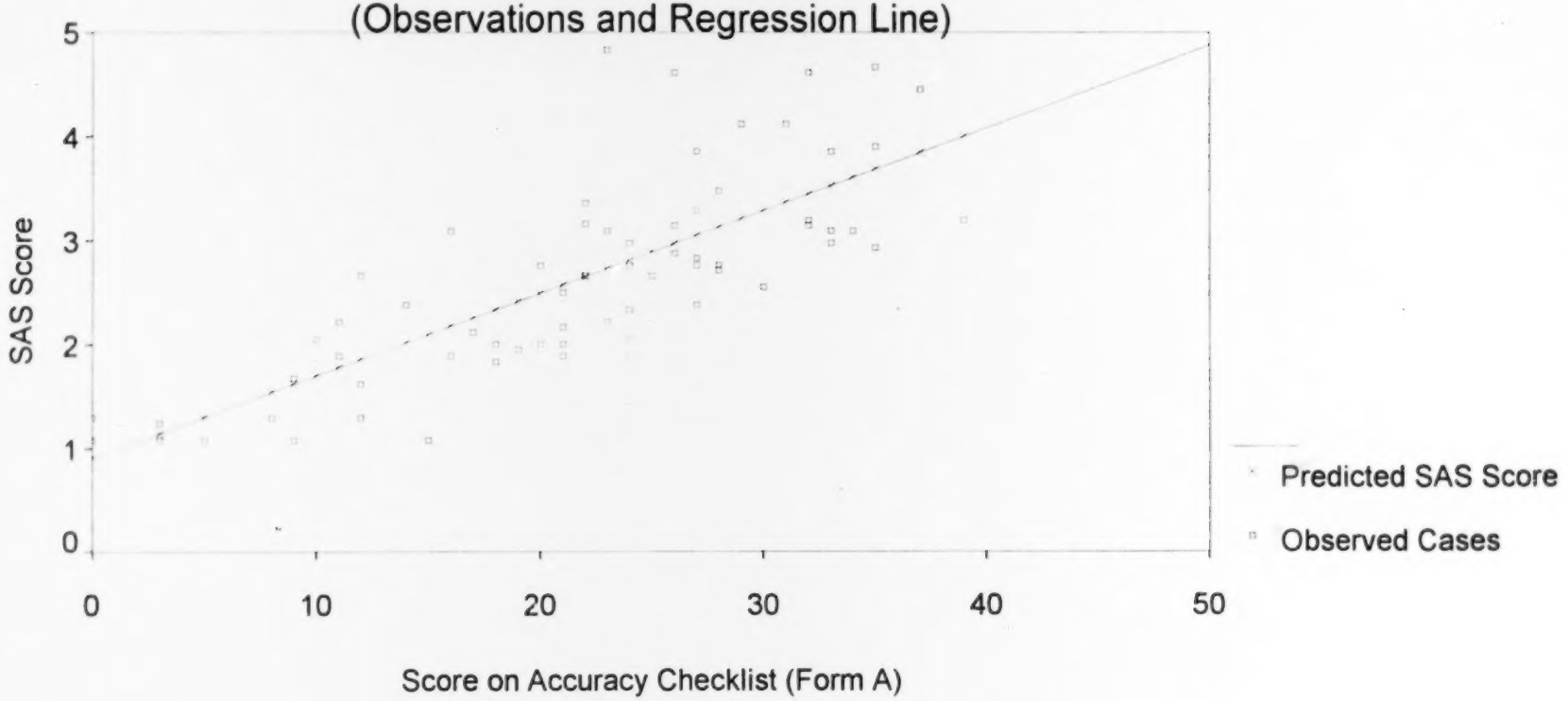


All Cases

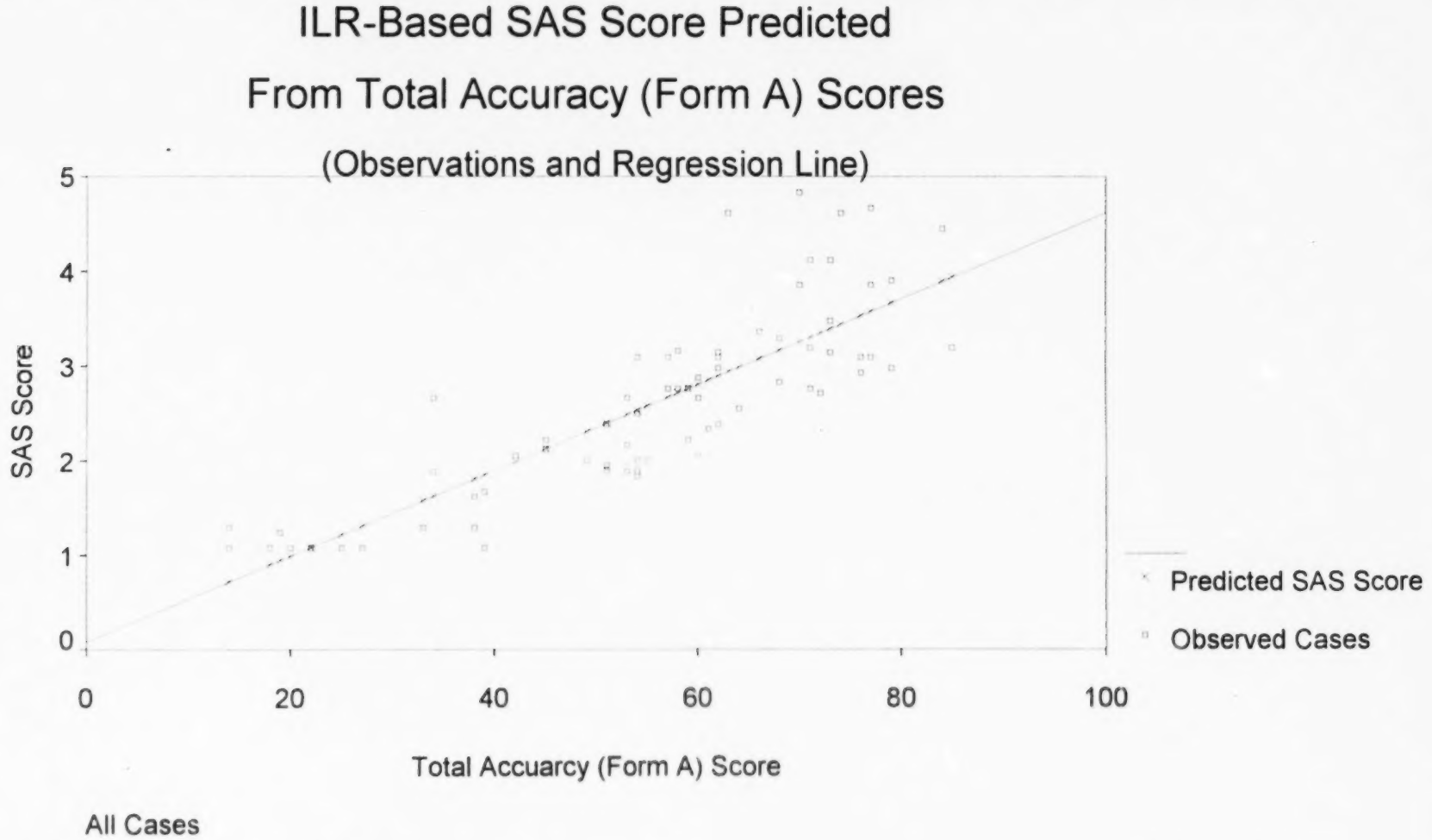
ILR-Based SAS Score

Predicted from Checklist Scores (Form A)

(Observations and Regression Line)



All Cases


ILR-Based SAS Score Predicted
From Total Accuracy (Form A) Scores
(Observations and Regression Line)
SAS Score
0
1
2
3
4
5
0
20
40
60
80
100
Total Accuarcy (Form A) Score
Predicted SAS Score
Observed Cases
All Cases

APPENDIX N

ABBREVIATIONS AND ABBREVIATION EQUIVALENCIES

Abbreviations and Abbreviation Equivalencies

In order to standardize the interpretation of the LSTE technical reports, this report has used abbreviations identical to or similar to those used in the LSTE-Spanish report. However, these standardized abbreviations were not employed at the time the database created and the statistical analysis was run. Because the statistical analyses for this study were turned in to the FBI, the list of abbreviations used in this report and their corresponding abbreviations used in the statistical analyses are enumerated here below. The abbreviation on the left refers to the abbreviation used in this report. It is identical or very similar to the abbreviation used in the LSTE-Spanish report. The abbreviation on the right is used in the statistical analysis printouts accompanying the deliverables.

SA-LC = SALISNUM. Self-Assessment of Minnan listening comprehension on an ILR type scale converted to a numerical value. Maximum score is 10 since 0+ was the lowest point on the scale.
SA-ST = SASTTOT. Self-Assessment of summary translation ability total score based on the total of 4 self-ratings using a 4 point scale for each rating. Maximum possible score is 16.
SA-EW = SAWRTNUM. Self-Assessment of English Writing ability on an ILR type scale converted to a numerical value. Maximum score is 10 since 0+ is the lowest point on the scale.
MCA = ASCORE. Score on Form A, Multiple-Choice section. Maximum score is 50.
MCB = BSCORE. Score on Form B, Multiple-choice section. Maximum score is 50.
ACCA = ACTOTAJJ. Accuracy checklist total, Form A, Rater 1 (Jing-Jing Liu). This is the sum of all points earned for messages conveyed on the three summaries on Form A when they are rated by rater 1.
ACCB = ACTOTBJJ. Accuracy checklist total, Form B, Rater 1 (Jing-Jing Liu)
TOTA = ACTOTALA. Accuracy total (MC + checklist) Form A. The Accuracy Total is the sum of correct answers on the multiple-choice and summary translation section.
TOTB = ACTOTALB. Accuracy total (MC + checklist) Form B
EXPA = EXAVALLA. Expression average (composite of ratings by two raters), Form A.
EXPB = EXAVALLB. Expression average (composite of ratings by two raters), Form B.
SAS = SELFILR. The Summary Accuracy Scale constructed for this study based on a scaling of the sum of the subject's SA-LC and SA-ST. SAS is reported on a 0-5 ILR-based scale.